UNIVERSIDAD NACIONAL AUTONOMA
DE MEXICO

POSGRADO EN CIENCIAS
MATEMATICAS
FACULTAD DE CIENCIAS

APLICACIONES BAYESIANAS DE MEZCLAS BASADAS EN PROCESOS DIRICHLET

T E S I S

QUE PARA OBTENER EL GRADO ACADEMICO DE

MAESTRO EN CIENCIAS

(MATEMATICAS)

P R E S E N T A

LUIS GONZALO LEON NOVELO

DIRECTOR DE TESIS:
DR. EDUARDO GUTIERREZ PEÑA

MEXICO, D. F. NOVIEMBRE 2005

Agradecimientos.

A mi mamá Enna:

Por darme lo más valioso que tengo...la vida.

A mis abuelos:

Enrique y Enna por ser mi ejemplo y compartir su vida conmigo. Josefa por quererme.

A mis hermanos:

Enna, Alejandro, Enrique y Oscar por ser mis mejores amigos. Los quiero.

A mis tíos y primos:

Graciela y Joaquín por consentirme y hacerme más agradable la vida y, a mis primos, Mariza y Joaquín por compatir su tiempo conmigo.

Al Dr. Eduardo Gutiérrez:

Por asesorarme, junto con Raúl, en la realización de esta tesis y darme el mejor panorama posible de lo que debía hacer. De verdad te lo agradezco mucho.

Al Dr. Raúl Rueda:

Por volver a confiar en mí, por tus enseñanzas, por guiarme y, una vez más, apoyarme y asesorarme en la realización de una tesis. Muchas gracias Raúl.

A mis sinodales:

Los Doctores Ramses Mena, Alberto Contreras y Carlos Díaz por la revisión de este trabajo y sus valiosas sugerencias.

A mis maestros:

En especial a Raúl Rueda, Eduardo Gutierréz, Federico O'Reilly, José María González, Alberto Contreras, Silvia Ruíz y Rafael Madrid por compartir su tiempo y conocimientos conmigo.

A mis amigos

Por su apoyo incondicional y por estar conmigo siempre.

Índice

1. INTRODUCCIÓN

Las distribuciones que se utilizan comúnmente en Estadística para modelar al proceso que genera los datos, tales como las distribuciones normal o Weibull, implican ciertos supuestos sobre las características de la población. En los casos mencionados, por ejemplo, hay un supuesto de unimodalidad y, por otra parte, no permiten modelar más allá de los dos primeros momentos poblacionales (media y varianza). Este tipo de modelos no son suficientes para analizar conjuntos de datos relativamente complejos.

Una posibilidad es buscar modelos paramétricos más flexibles, pero en general estos modelos seguirán imponiendo ciertas restricciones sobre las características de la población. Otra alternativa es utilizar mezclas finitas de distribuciones, las cuales producen clases de modelos bastante más flexibles, aunque su análisis es considerablemente más complicado. Finalmente, podría considerarse el uso de modelos con un número infinito de parámetros. Los modelos Bayesianos no paramétricos pertenecen a esta clase.

Como su nombre lo indica, la estadística Bayesiana no paramétrica surge del deseo de no suponer una familia paramétrica de distribuciones específica para la modelación de los datos, lo que permite una descripción más realista de la incertidumbre sobre el proceso que los generó, así como inferencias más robustas que las obtenidas con modelos paramétricos.

En el análisis Bayesiano tradicional, se asume que la densidad de las observaciones, $p(y|\theta)$, y la del parámetro $p(\theta)$ son conocidas. A partir de ellas es posible hallar la distribución final de θ, $p(\theta|y)$, así como la distribución predictiva de una observación futura y_f, $p(y_f|y)$, que son la base para cualquier tipo de inferencia.

Para cada pareja diferente de $p(y|\theta)$ y $p(\theta)$ que supongamos llegamos a distintos resultados. Si bien muchas veces se elige $p(\theta)$ de forma que los cálculos sean sencillos, o que $p(\theta)$ sea no informativa, al final $p(\theta)$ no influirá significativamente sobre los resultados siempre que el tamaño de la muestra sea suficientemente grande. Sin embargo, la elección de $p(y|\theta)$ sí puede cambiar en gran medida los resultados. Con la finalidad de no suponer ninguna distribución para $p(y|\theta)$ podemos pensar a θ como una función de densidad (o de distribución), y para ser más explícitos en vez de $p(y|\theta)$ escribiremos $p(y|f)$ (respectivamente $p(y|F)$) donde f (F) es una función de densidad (distribución), y de esta forma $y|f$ ($y|F$) tiene la densidad (distribución) f (F). De esta manera el problema se complica, pues si queremos resolver el problema con la teoría estadística Bayesiana requerimos de

una distribución inicial Π para el parámetro f (F), es decir, necesitamos de una medida de probabilidad sobre el espacio de funciones de densidad (distribución).

Por otro lado, notemos que si Π, ahora pensándola como medida de probabilidad, da peso 1 a una familia parametrizada por algún parámetro τ (por ejemplo a la familia de normales con media τ y varianza 1), entonces la medida Π es equivalente a una medida π en el espacio de posibles valores de τ, es decir,

$$Pr[f \in A] = \Pi(A) = \int_{\{\tau: f(\cdot,\tau) \in A\}} \pi(\tau)\, d\tau.$$

Por lo tanto, si pensamos a f como un parámetro entonces podemos englobar a la estadística no paramétrica en la paramétrica, y, por otro lado, si Π da probabilidad 1 a una familia paramétrica, entonces el enfoque no paramétrico coincide con el paramétrico. Entenderemos que se está resolviendo un problema estadístico desde el enfoque de la estadística Bayesiana no paramétrica siempre que Π no de probabilidad 1 a una familia paramétrica cuyos parámetros tengan dimensión finita.

Encontrar una medida de probabilidad Π manejable con estas características no fue sencillo. Las primeras distribuciones iniciales para problemas no paramétricos fueron estudiadas por Freedman (1963) quien introdujo las medidas aleatorias de Dirichlet. Posteriormente, Freedman (1965), Dubins y Freedman (1967) y Ferguson (1973,1974) formalizaron y exploraron con mayor detalle al proceso Dirichltet. Algunas de las generalizaciones más importantes incluyen las mezclas de procesos Dirichlet (Antoniak, 1974), el proceso estable normalizado (Kingman, 1975) y los árboles de Pólya (Lavine, 1992, 1994). Sin embargo, la utilización de la medida de Ferguson, conocida como proceso Dirichlet no fue posible sino hasta que el desarrollo de las computadoras y de los métodos de simulación, como el de Monte Carlo, lo permitieron.

Los modelos y métodos Bayesianos no paramétricos han alcanzado la madurez suficiente como para ofrecer una alternativa viable para el análisis de datos en una gran variedad de aplicaciones. Dey et al. (1998) y Walker et al. (1999) discuten e ilustran algunos de los modelos no paramétricos y semiparamétricos disponibles en la actualidad.

La ventaja del proceso Dirichlet es la facilidad con la que se encuentra la distribución final de la función de distribución desconocida dada una realización de la muestra, pues dicha distribución es también un proceso Dirichlet. La desventaja es que asigna proba-

bilidad 1 a las distribuciones discretas. Aunque en la actualidad existen diversas clases de distribuciones iniciales Π con soporte en las distribuciones continuas, el modelo más utilizado es el de las mezclas basadas en procesos Dirichlet. La versión actual de esta clase de modelos se basa en los trabajos de Antoniak (1974), Lo (1984) y Escobar (1988).

Esta tesis aborda la definición de proceso Dirichlet de Ferguson y la de mezclas de procesos Dirichlet de Antoniak en el segundo capítulo, las mezclas basadas en procesos Dirichlet propuestas por Escobar y algunas de sus aplicaciones en el tercero y, finalmente, en el cuarto se da una aplicación de las mezclas basadas en procesos Dirichlet en el problema de selección de modelos paramétricos desde el punto de vista Bayesiano no paramétrico.

2. PROCESO DIRICHLET Y MEZCLAS DE PROCESOS DIRICHLET

2.1 *Preliminares*

En ciertos problemas estadísticos, sólo se sabe que la distribución F en el espacio muestral Θ pertenece a algún conjunto de distribuciones $\mathcal{F} = \{F_\alpha\}$. Este conjunto $\mathcal{F}$ puede ser tratado como un espacio parametral en un análisis Bayesiano de estos problemas. Si suponemos que $\mathcal{F}$ es una familia de distribuciones paramétrica, es decir, está caracterizada por un número finito de parámetros, podemos dar la distribución inicial sobre los parámetros. Por otro lado, si $\mathcal{F}$ no es una familia paramétrica, por ejemplo la clase de todas las funciones de distribución continuas en la recta real, estamos suponiendo un modelo no paramétrico y debemos asignar una distribución inicial sobre $\mathcal{F}$.

De acuerdo con Antoniak (1974), las propiedades que desearíamos que satisfaciera una familia de distribuciones iniciales $\mathcal{D}$ sobre $\mathcal{F}$ son las siguientes:

1. Qué $\mathcal{D}$ sea matemáticamente tratable en tres aspectos:
   (a) La distribución final en $\mathcal{F}$ dada una muestra, debe ser razonablemente sencilla;
   (b) Debe ser posible expresar convenientemente la esperanza de una función de pérdida simple; y
   (c) $\mathcal{D}$ debe ser cerrada, en el sentido de que si la inicial está en $\mathcal{D}$, entonces la final también.
2. Qué $\mathcal{D}$ sea "rica", es decir, debe existir un elemento en $\mathcal{D}$ capaz de expresar cualquier información inicial.
3. Qué $\mathcal{D}$ sea parametrizada de manera que los parámetros puedan ser interpretados en relación a la información que se tiene y al grado de credibilidad que se tiene en ella.

Estas propiedades no son mutuamente excluyentes, pero sí antagónicas en el sentido de que se puede obtener una a expensas de las otras. Por ejemplo, algunos autores, entre ellos Dubins y Freedman (1965,1967), Kraft y Van Eeden (1964) y Kraft (1964) han descrito familias $\mathcal{D}$ que satisfacen la segunda propiedad pero son deficientes en la primera y la tercera.

En este trabajo abordaremos el proceso estocástico definido por Ferguson (1973) llamado proceso Dirichlet, que es particularmente eficiente en la primera y tercera propiedades, pero no en la segunda. Para comprender el proceso Dirichlet, veamos primero algunas

propiedades de la distribución Dirichlet.

2.2 *Distribución Dirichlet*

La distribución Dirichlet es conocida como la distribución conjugada para los parámetros de una distribución multinomial.

Denotemos por $\mathcal{G}(\cdot|\alpha,\beta)$ a la distribución gamma con parámetro de forma $\alpha \geqslant 0$ y de escala $\beta > 0$. Para $\alpha = 0$, esta distribución es degenerada en cero; para $\alpha > 0$, esta distribución tiene densidad respecto a la medida de Lebesgue en la recta real dada por

$$f(z|\alpha,\beta) = \frac{1}{\Gamma(\alpha)}\beta^{\alpha}\exp(-z\beta)z^{\alpha-1}1_{(0,\infty)}(z), \tag{2.1}$$

donde $1_S(z)$ denota a la función indicadora del conjunto S.

Generalmente se define la distribución Dirichlet con parámetros positivos de acuerdo a Wilks (1962), nosotros utilizaremos la extensión de Ferguson (1973) permitiendo que algunos, pero no todos, sus parámetros sean iguales a cero.

Sean $Z_1,\ldots,Z_k$ variables aleatorias independientes con distribución $\mathcal{G}(\alpha_j,1)$ respectivamente, con $\alpha_j \geqslant 0$ para toda j y $\alpha_j > 0$ para alguna j, $j = 1,\ldots,k$. La distribución Dirichlet con parámetros $\alpha_1,\ldots,\alpha_k$, denotada por $\mathcal{D}(\alpha_1,\ldots,\alpha_k)$, está definida como la distribución de $(Y_1,\ldots,Y_k)$, donde

$$Y_j = Z_j \Big/ \sum_{i=1}^{k} Z_i \qquad \text{para } j = 1,\ldots,k. \tag{2.2}$$

Al usar la notación $\mathcal{D}(\alpha_1,\ldots,\alpha_k)$ entenderemos que $\alpha_j \geqslant 0$ para toda j, y $\alpha_j > 0$ para alguna j. Esta distribución es siempre singular con respecto a la medida de Lebesgue en $\mathbb{R}^k$ ya que $Y_1 + \ldots + Y_k = 1$. Además, si $\alpha_j = 0$, la correspondiente Y_j es degenerada en 0. Sin embargo, si $\alpha_j > 0$ para toda j, la distribución en $\mathbb{R}^{k-1}$ de $(Y_1,\ldots,Y_{k-1})$ es absolutamente continua con respecto a la medida de Lebesgue en $\mathbb{R}^{k-1}$ y tiene como función de densidad†

$$f(y_1,\ldots,y_{k-1}|\alpha_1,\ldots,\alpha_k) = \frac{\Gamma(\alpha_1+\ldots+\alpha_k)}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_k)}\prod_{j=1}^{k-1} y_j^{\alpha_j-1}\left(1-\sum_{i=1}^{k-1} y_i\right)^{\alpha_k-1} 1_S(y_1,\ldots,y_{k-1}), \tag{2.3}$$

† Ver Wilks (1962), p. 179

donde S es el conjunto

$$S = \left\{ (y_1, \ldots, y_{k-1}) \in \mathbb{R}^{k-1} : \; y_j \geqslant 0, \; \sum_{j=1}^{k-1} y_j \leqslant 1 \right\}.$$

Para $k = 2$, (2.3) se reduce a la distribución Beta, denotada por $\mathcal{B}e(\cdot|\alpha_1, \alpha_2)$.

Una propiedad importante de la distribución Dirichlet que, como se verá más adelante, asegura la existencia del proceso Dirichlet es la siguiente:

i°. Si $(Y_1, \ldots, Y_k) \sim \mathcal{D}(\alpha_1, \ldots, \alpha_k)$ y $r_1, \ldots, r_l$ son enteros tales que $0 < r_1 < \ldots < r_l = k$, entonces,

$$\left(\sum_{i=1}^{r_1} Y_i, \sum_{i=r_1+1}^{r_2} Y_i, \ldots, \sum_{i=r_{l-1}+1}^{r_l} Y_i \right) \sim \mathcal{D}\left(\sum_{i=1}^{r_1} \alpha_i, \sum_{i=r_1+1}^{r_2} \alpha_i, \ldots, \sum_{i=r_{l-1}+1}^{r_l} \alpha_i \right).$$

Esta propiedad, que es válida incluso si $\alpha_j = 0$ para alguna j, se sigue directamente de la definición y la propiedad aditiva de la distribución gamma: si $Z_1 \sim \mathcal{G}(\alpha_1, 1)$, $Z_2 \sim \mathcal{G}(\alpha_2, 1)$ y son independientes entonces $Z_1 + Z_2 \sim \mathcal{G}(\alpha_1 + \alpha_2, 1)$. En particular, la distribución marginal de cada Y_j es Beta: $Y_j \sim \mathcal{B}e(\alpha_j, \sum_{i=1}^{k} \alpha_i - \alpha_j)$.

Utilizaremos más adelante los primeros dos momentos de la distribución Dirichlet.

ii°. Si $(Y_1, \ldots, Y_k) \sim \mathcal{D}(\alpha_1, \ldots, \alpha_k)$, entonces

$$\begin{aligned} E(Y_i) &= \alpha_i/\alpha, \\ E(Y_i^2) &= \alpha_i(\alpha_i + 1)/(\alpha(\alpha + 1)) \qquad \text{y} \\ E(Y_i Y_j) &= \alpha_i \alpha_j/(\alpha(\alpha + 1)), \qquad \text{para } i \neq j, \end{aligned}$$

donde $\alpha = \sum_{i=1}^{k} \alpha_i$.

La siguiente propiedad también es conocida.

iii°. Si la distribución inicial de $(Y_1, \ldots, Y_k)$ es $\mathcal{D}(\alpha_1, \ldots, \alpha_k)$ y si θ es una variable aleatoria tal que,

$$\mathbb{P}\{\theta = j | Y_1, \ldots, Y_k\} = Y_j \qquad c.s., \qquad \text{para } j = 1, \ldots, k,$$

entonces la distribución final de $(Y_1, \ldots, Y_k)$ dado $\theta = j$ es $\mathcal{D}(\alpha_1^{(j)}, \ldots, \alpha_k^{(j)})$, donde

$$\alpha_i^{(j)} = \begin{cases} \alpha_i, & \text{si } i \neq j \\ \alpha_i + 1, & \text{si } i = j. \end{cases}$$

Usaremos $\mathcal{D}(y_1,\ldots,y_k|\alpha_1,\ldots,\alpha_k)$ para denotar la función de distribución Dirichlet evaluada en $(y_1,\ldots,y_k)$. Utilizando ii^o y iii^o obtenemos

$$\begin{aligned}
&\int_0^{z_1}\cdots\int_0^{z_k} y_j\, d\mathcal{D}(y_1,\ldots,y_k|\alpha_1,\ldots,\alpha_k)\\
&\quad=\int_0^{z_1}\cdots\int_0^{z_k}\mathbb{P}\{\theta=j|Y_1=y_1,\ldots,Y_k=y_k\}\, d\mathcal{D}(y_1,\ldots,y_k|\alpha_1,\ldots,\alpha_k)\\
&\quad=\mathbb{P}\{\theta=j,Y_1\leqslant z_1,\ldots,Y_k\leqslant z_k\}\\
&\quad=\mathbb{P}\{\theta=j\}\mathbb{P}\{Y_1\leqslant z_1,\ldots,Y_k\leqslant z_k|\theta=j\}\\
&\quad=\int_0^1\cdots\int_0^1\mathbb{P}\{\theta=j|Y_1=y_1,\ldots,Y_k=y_k\}\, d\mathcal{D}(y_1,\ldots,y_k|\alpha_1,\ldots,\alpha_k) \qquad (2.4)\\
&\qquad\times\mathcal{D}(z_1,\ldots,z_k|\alpha_1^{(j)},\ldots,\alpha_k^{(j)})\\
&\quad=\int_0^1\cdots\int_0^1 y_j\, d\mathcal{D}(y_1,\ldots,y_k|\alpha_1,\ldots,\alpha_k)\;\mathcal{D}(z_1,\ldots,z_k|\alpha_1^{(j)},\ldots,\alpha_k^{(j)})\\
&\quad=E(Y_j)\mathcal{D}(z_1,\ldots,z_k|\alpha_1^{(j)},\ldots,\alpha_k^{(j)})\\
&\quad=\frac{\alpha_j}{\alpha}\mathcal{D}(z_1,\ldots,z_k|\alpha_1^{(j)},\ldots,\alpha_k^{(j)}).
\end{aligned}$$

Debemos notar que esta expresión es válida aún si $\alpha_j=0$.

2.3 *Proceso Dirichlet*

Sea $(\Theta,\mathcal{A})$ un espacio de medida, el objetivo es encontrar una medida de probabilidad en el espacio de medidas de probabilidad de $(\Theta,\mathcal{A})$. Por ejemplo, si $\Theta=\{1,2,\ldots,k\}$ es finito, cualquier medida de probabilidad p está determinada por el vector $(p_1=p(\{1\}),p_2=p(\{2\}),\ldots,p_k=p(\{k\}))$ tomando valores en el conjunto $\Delta_k=\{(p_1,p_2,\ldots,p_k):0\leqslant p_i\leqslant 1,\ \sum p_i=1,1\leqslant i\leqslant k\}$ y una manera de probabilizar dicho espacio es asignando una distribución Dirichlet al vector $(p_1,p_2,\ldots,p_k)$. Esta idea se generaliza para el caso infinito a partir del proceso Dirichlet, cuya construcción desarrollamos a continuación.

Sea $(\Theta,\mathcal{A})$ un espacio de medida y definamos

$$[0,1]^{\mathcal{A}}=\{f:\mathcal{A}\to[0,1]\}$$

y $\mathcal{BF}$ la $\sigma-$álgebra generada por los conjuntos de $[0,1]^{\mathcal{A}}$ de la forma

$$\{f\in[0,1]^{\mathcal{A}}:f(A)\leqslant t\},t\in[0,1],A\in\mathcal{A}.$$

La idea es probar la existencia de una medida de probabilidad $\mathbb{P}$ en $([0,1]^{\mathcal{A}},\mathcal{BF})$ (que en el caso $(\Theta,\mathcal{A})=(\mathbb{R},\mathcal{B}(\mathbb{R}))$ de peso 1 al conjunto de todas las medidas de probabilidad en $\mathbb{R}$).

Para cada A en $\mathcal{A}$ se define $P_A : [0,1]^{\mathcal{A}} \to [0,1]$ como la proyección natural, es decir, dado f en $[0,1]^{\mathcal{A}}$

$$P_A f = f(A),$$

notando que P_A es $\mathcal{BF}$-medible para todo A en $\mathcal{A}$ tenemos que $\mathcal{P} = \{P_A : A \in \mathcal{A}\}$ es un proceso estocástico con valores en $[0,1]$ y conjunto de índices $\mathcal{A}$.

Una vez especificada la distribución conjunta de las variables aleatorias $P_{B_1}, \ldots, P_{B_k}$ para toda $k \in \mathbb{N}$ y toda partición $\{B_1, \ldots, B_k\}$ medible de Θ ($\{B_1, \ldots, B_k\}$ es una partición medible de Θ, si $B_i \in \mathcal{A}$ para $i \in \{0, 1, \ldots, k\}$, $B_i \cap B_j = \emptyset$ para $i \neq j$ y $\bigcup_{i=1}^{k} B_i = \Theta$), podemos definir la distribución conjunta de $P_{A_1}, \ldots, P_{A_m}$ para $m \in \mathbb{N}$ y $A_1, A_2, \ldots, A_m \in \mathcal{A}$ si degeneramos a $P_\emptyset$ en cero, es decir, $P_\emptyset = 0$ con probabilidad 1, de la siguiente manera:

Dados $A_1, A_2, \ldots A_m \in \mathcal{A}$, formemos los $k = 2^m$ conjuntos obtenidos tomando las intersecciones de los $A_i's$ y sus complementos; esto es, definiendo $B(\nu_1, \nu_2, \ldots, \nu_m)$ para $\nu_j = 0$ ó 1 de la siguiente forma

$$B(\nu_1, \nu_2, \ldots, \nu_m) = \bigcap_{j=1}^{m} A_j^{\nu_j}, \tag{2.5}$$

donde A_j^1 está definida como A_j y A_j^0 como el complemento de A_j. Observemos que,

$$A_j = \bigcup \{B(\nu_1, \nu_2, \ldots, \nu_m) : \nu_i \in \{0,1\}, i = 1, \ldots, m \text{ y } \nu_j = 1\}, \qquad \text{para } j = 1, \ldots, m,$$

y como $\{B(\nu_1, \nu_2, \ldots, \nu_m) : \nu_i \in \{0,1\}, i \in \{1, \ldots, n\}\}$ forma una partición medible de Θ, suponemos que conocemos la distribución conjunta de

$$\left\{P_{B(\nu_1,\nu_2,\ldots,\nu_m)} : \nu_i \in \{0,1\}, i \in \{1, \ldots, m\}\right\}, \tag{2.6}$$

y a partir de ella definimos a la distribución conjunta de $(P_{A_1}, P_{A_2}, \ldots, P_{A_m})$ como la del vector

$$\Big(\sum_{\{\nu_1,\ldots,\nu_m : \nu_1 = 1\}} P_{B(\nu_1,\nu_2,\ldots,\nu_m)}, \sum_{\{\nu_1,\ldots,\nu_m : \nu_2 = 1\}} P_{B(\nu_1,\nu_2,\ldots,\nu_m)}, \ldots, \sum_{\{\nu_1,\ldots,\nu_m : \nu_m = 1\}} P_{B(\nu_1,\nu_2,\ldots,\nu_m)} \Big). \tag{2.7}$$

Notemos que con esta definición no hay contradicción si $B(\nu_1, \nu_2, \ldots, \nu_m) = \emptyset$ ya que pedimos que $P_\emptyset = 0$ con probabilidad 1.

Una condición que nos asegura que un sistema de distribuciones para $(P_{B_1}, \ldots, P_{B_k})$ para toda $k \in \mathbb{N}$ y toda partición medible $B_1, B_2, \ldots, B_k$ satisface la segunda condición de consistencia de Kolmogorov (1933) es la siguiente:

Condición (C).

Si $\{B'_1, B'_2, \ldots, B'_{k'}\}$ es un refinamiento de la partición $\{B_1, B_2, \ldots, B_k\}$, con $B_1 = \cup_{i=1}^{r_1} B'_i$, $B_2 = \cup_{i=r_1+1}^{r_2} B'_i, \ldots, B_k = \cup_{i=r_{k-1}+1}^{r_k} B'_i$, con $1 = r_1 < r_2 < \ldots < r_k = k'$, entonces la distribución de

$$\left(\sum_{i=1}^{r_1} P_{B'_i}, \sum_{i=r_1+1}^{r_2} P_{B'_i}, \ldots, \sum_{i=r_{k-1}+1}^{r_k} P_{B'_i} \right)$$

determinada por la distribución conjunta de $(P_{B'_1}, P_{B'_2}, \ldots, P_{B'_k})$ es la misma que la de $(P_{B_1}, P_{B_2}, \ldots, P_{B_k})$.

Definición 2.1 *Decimos que el proceso estocástico* $\{Q_A : A \in \mathcal{A}\}$ *es una medida de probabilidad aleatoria finitamente aditiva en* $(\Theta, \mathcal{A})$, *si satisface la condición* (C), Q_A *toma valores en* $[0,1]$ *y* $Q_\Theta = 1$ *con probabilidad* 1.

Notemos que si Q es una medida de probabilidad aleatoria finitamente aditiva y $A_1, \ldots, A_m \in \mathcal{A}$ son ajenos, entonces $A_1, \ldots, A_m, A_{m+1}$ y $\cup_{i=1}^m A_i, A_{m+1}$ son dos particiones de Θ con $A_{m+1} \equiv \left(\cup_{i=1}^m A_i\right)^c$, por lo que $\sum_{i=1}^m Q_{A_i} + Q_{A_{m+1}}$ y $Q_{\cup_{i=1}^m A_i} + Q_{A_{m+1}}$ tienen, por la condición (C), la misma distribución que $Q_\Theta = 1$ *c.s.*, de ahí que $Q_{\cup_{i=1}^m A_i} = \sum_{i=1}^m Q_{A_i}$ *c.s.*

Un ejemplo de una medida de probabilidad aleatoria finitamente aditiva en $(\Theta, \mathcal{A})$ es: dadas $X_1, \ldots, X_n$ variables aleatorias con valores en Θ, definimos $Q = \{Q_{n,A} : A \in \mathcal{A}\}$ como

$$Q_{n,A} = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}(A),$$

donde δ_x denota la delta de Dirac, que es la medida que da masa unitaria al punto x, es decir,

$$\delta_x(A) = \begin{cases} 1, & \text{si } x \in A \\ 0, & \text{si } x \notin A. \end{cases}$$

Definición 2.2 *Un sistema de distribuciones para el proceso estocástico* $\{y_t : t \in T\}$ *con valores en* E *y conjunto de índices* T *es libre de orden si satisface la primera condición de consistencia de Kolmogorov (1933), es decir: para toda* $n \in \mathbb{N}$, $H_1, \ldots, H_n$ *en la* σ*-álgebra de* E, $t_1, \ldots, t_n \in T$ *y toda permutación* π *de* $\{1, \ldots, n\}$,

la probabilidad que asigna el vector aleatorio $(y_{t_1} \ldots, y_{t_n})$ *al conjunto* $H_1 \times \ldots \times H_n$ *es igual a la que asigna* $(y_{t_{\pi(1)}} \ldots, y_{t_{\pi(n)}})$ *a* $H_{\pi(1)} \times \ldots \times H_{\pi(n)}$.

LEMA 2.1 *Si un sistema de distribuciones para* $(P_{B_1}, P_{B_2}, \ldots, P_{B_k})$ *para toda* $k \in \mathbb{N}$ *y toda* $\{B_1, B_2, \ldots, B_k\}$ *partición medible de* Θ *es libre de orden, satisface la condición* (C), *y si para cualesquiera conjuntos* $A_1, A_2, \ldots, A_m \in \mathcal{A}$, *la distribución* $(P_{A_1}, P_{A_2}, \ldots, P_{A_m})$ *es definida como en* (2.5), (2.6), *y* (2.7) *entonces existe una medida de probabilidad* $\mathbb{P}$ *en* $([0,1]^{\mathcal{A}}, \mathcal{BF})$ *cuyas distribuciones finito dimensionales coinciden con las de este sistema.*

DEMOSTRACIÓN. Como $\Theta \cup \emptyset = \Theta$, se sigue de la condición (C) que $P_\emptyset = 0$ con probabilidad 1; por lo tanto la distribución de $(P_{A_1}, P_{A_2}, \ldots, P_{A_m})$ está bien definida. Para demostrar el lema debemos verificar la primera y segunda condiciones de consistencia de Kolmogorov (1933, p. 29). La primera es automática, pues estamos suponiendo que la distribución $(P_{B_1}, P_{B_2}, \ldots, P_{B_k})$ es libre de orden.

Para la segunda, debemos probar que para m arbitraria y $A_1, A_2, \ldots, A_m \in \mathcal{A}$, la distribución marginal de $(P_{A_1}, P_{A_2}, \ldots, P_{A_{m-1}})$ derivada de la de $(P_{A_1}, P_{A_2}, \ldots, P_{A_m})$ es idéntica a la distribución de $(P_{A_1}, P_{A_2}, \ldots, P_{A_{m-1}})$ definida directamente.

La distribución marginal de $(P_{A_1}, P_{A_2}, \ldots, P_{A_{m-1}})$ derivada de la distribución marginal de $(P_{A_1}, P_{A_2}, \ldots, P_{A_m})$ es idéntica a la distribución de

$$\Big(\sum_{\{\nu_1,\ldots,\nu_m:\nu_1=1\}} P_{B(\nu_1,\nu_2,\ldots,\nu_m)}, \sum_{\{\nu_1,\ldots,\nu_m:\nu_2=1\}} P_{B(\nu_1,\nu_2,\ldots,\nu_m)}, \ldots, \sum_{\{\nu_1,\ldots,\nu_m:\nu_{m-1}=1\}} P_{B(\nu_1,\nu_2,\ldots,\nu_m)} \Big) \quad (2.8)$$

derivada de (2.6) y la distribución de $(P_{A_1}, P_{A_2}, \ldots, P_{A_{m-1}})$ está definida como la distribución de

$$\Big(\sum_{\{\nu_1,\ldots,\nu_{m-1}:\nu_1=1\}} P_{B(\nu_1,\nu_2,\ldots,\nu_{m-1})}, \sum_{\{\nu_1,\ldots,\nu_{m-1}:\nu_2=1\}} P_{B(\nu_1,\nu_2,\ldots,\nu_{m-1})}, \ldots, \sum_{\{\nu_1,\ldots,\nu_{m-1}:\nu_{m-1}=1\}} P_{B(\nu_1,\nu_2,\ldots,\nu_{m-1})} \Big). \quad (2.9)$$

derivada de la distribución de $\{P_{B(\nu_1,\nu_2,\ldots,\nu_{m-1})} : \nu_i \in \{0,1\}, i \in \{1,\ldots,m-1\}\}$, donde

$$B(\nu_1, \nu_2, \ldots, \nu_{m-1}) = \bigcap_{j=1}^{m-1} A_j^{\nu_j},$$

como $B(\nu_1, \nu_2, \ldots, \nu_{m-1}) = B(\nu_1, \nu_2, \ldots, \nu_{m-1}, 0) \cup B(\nu_1, \nu_2, \ldots, \nu_{m-1}, 1)$. La condición (C) implica que la distribución de $\{P_{B(\nu_1,\nu_2,\ldots,\nu_{m-1})} : \nu_i \in \{0,1\}, i \in \{1,\ldots,m-1\}\}$ es idéntica

a la de

$$\{P_{B(\nu_1,\nu_2,\ldots,\nu_{m-1},0)} + P_{B(\nu_1,\nu_2,\ldots,\nu_{m-1},1)} : \nu_i \in \{0,1\}, i \in \{1,\ldots,m-1\}\}$$

determinada por la distribución de (2.6) por lo tanto la distribución de (2.9) puede ser encontrada también de la distribución de (2.6) remplazando $P_{B(\nu_1,\nu_2,\ldots,\nu_{m-1})}$ por

$$P_{B(\nu_1,\nu_2,\ldots,\nu_{m-1},0)} + P_{B(\nu_1,\nu_2,\ldots,\nu_{m-1},1)}.$$

Con este remplazo (2.9) es formalmente idéntico a (2.8), probando que sus distribuciones son idénticas. Con lo que queda demostrado el lema.

◁

Definición 2.3 *Sea α una medida no nula y finitamente aditiva en $(\Theta, \mathcal{A})$. Decimos que $\mathcal{P} = \{P_A : A \in \mathcal{A}\}$ es un proceso Dirichlet con parámetro α, y lo denotamos por $\mathcal{P} = \mathcal{DP}(\alpha)$, si para todo k en $\mathbb{N}$ y para toda partición medible $\{B_1, B_2, \ldots, B_k\}$ de Θ, la distribución de $(P_{B_1}, P_{B_2}, \ldots, P_{B_k})$ es Dirichlet, $\mathcal{D}(\alpha(B_1), \alpha(B_2), \ldots, \alpha(B_k))$.*

La condición de consistencia (C) para el proceso Dirichlet es exactamente la propiedad i^o de la distribución Dirichlet, que es obviamente una distribución libre de orden. Se sigue del Lema 2.1, la existencia de una medida de probabilidad $\mathbb{P}$ en $([0,1]^{\mathcal{A}}, \mathcal{BF})$ tal que las proyecciones $\mathcal{P} = \{P(A) : A \in \mathcal{A}\}$ bajo $\mathbb{P}$ es un proceso Dirichlet, *i.e.*,

$$\mathbb{P}[P_{B_1} \leqslant t_1, P_{B_2} \leqslant t_2, \ldots, P_{B_k} \leqslant t_k] = \mathcal{D}(t_1, t_2, \ldots, t_k | \alpha(B_1), \alpha(B_2), \ldots, \alpha(B_k))$$

Además como P_Θ es degenerada en 1, por la definición de la distribución Dirichlet, $\mathcal{P}$ es una medida de probabilidad aleatoria finitamente aditiva. Las siguientes tres proposiciones muestran la relación entre las propiedades de la medida de probabilidad $\mathbb{P}$ y las propiedades del parámetro del proceso, α.

Proposición 2.1 *Sea $\mathcal{P}$ un proceso Dirichlet en $(\Theta, \mathcal{A})$ con parámetro α, y $A \in \mathcal{A}$. Si $\alpha(A) = 0$, entonces $P_A = 0$ con probabilidad 1 y si $\alpha(A) > 0$, entonces $P_A > 0$ con probabilidad 1. Además $E(P_A) = \alpha(A)/\alpha(\Theta)$.*

Demostración. Considerando la partición A, A^c, notamos que P_A tiene una distribución $\mathcal{B}e(\alpha(A), \alpha(A^c))$, utilizando la propiedad ii^o de la distribución Dirichlet obtenemos el

resultado.

◁

PROPOSICIÓN 2.2 *Sea $\mathcal{P}$ un proceso Dirichlet en $(\Theta, \mathcal{A})$ con parámetro α, si α es σ-aditiva, entonces también lo es $\mathcal{P}$, en el siguiente sentido: si $A_1, A_2, \ldots \in \mathcal{A}$ y $A_n \searrow \emptyset$, entonces $P_{A_n} \longrightarrow 0$ con probabilidad 1.*

DEMOSTRACIÓN. Supongamos que $\{A_n : n \in \mathbb{N}\} \searrow \emptyset$, entonces como α es σ-aditiva $\alpha(A_n) \to 0$. Por lo tanto existe una subsucesión $\{A_{n_j}\}$ tal que $\sum_{j=1}^{\infty} \alpha(A_{n_j}) < \infty$. Para ϵ fija, por la desigualdad de Tchebysheff,

$$\sum_{j=1}^{\infty} \mathbb{P}[P_{A_{n_j}} > \epsilon] \leqslant \frac{1}{\epsilon} \sum_{j=1}^{\infty} E[P_{A_{n_j}}] = \frac{1}{\epsilon} \sum_{j=1}^{\infty} \frac{\alpha(A_{n_j})}{\alpha(\Theta)} < \infty,$$

entonces por el teorema de Borel-Cantelli †, $\mathbb{P}[\limsup[P_{A_{n_j}} > \epsilon]] = 0$, es decir,

$$\mathbb{P}[P_{A_{n_j}} > \epsilon, \text{ para una infinidad de } j\text{'s}] = 0,$$

de lo que se tiene que

$$\mathbb{P}[P_{A_{n_j}} > \epsilon, \text{ en a lo más un número finito de } j\text{'s}] = 1,$$

de ahí que $P_{A_{n_j}} \to 0$ con probabilidad 1. Observando que $P_{A_n} \geqslant P_{A_{n+1}}$ $\mathbb{P}$-*c.s.* (ya que, por la condición (C), $P^c_{A_n} + P_{A_n \setminus A_{n+1}} + P_{A_{n+1}} = P^c_{A_n} + P_{A_n}$ $\mathbb{P}$-*c.s.*) se tiene el resultado. El inverso también es cierto, si α no es σ-aditiva $\mathcal{P}$ tampoco lo es con probabilidad 1.

◁

De la Proposición 2.2, se tiene que si α es σ-aditiva y $A, A_1, A_2, \ldots \in \mathcal{A}$ y $A_n \searrow A$ entonces,

$$P_{A_n} \to P_A \quad \mathbb{P} - c.s., \tag{2.10}$$

ya que $A \subset A_n \Rightarrow A \cup (A_n \setminus A) = A_n \Rightarrow P_A + P_{A_n \setminus A} = P_{A_n}$ $\mathbb{P}$-*c.s.* $\Rightarrow P_{A_n \setminus A} = P_{A_n} - P_A$ $\mathbb{P}$-*c.s.*, y como $A_n \setminus A \searrow \emptyset \Rightarrow P_{A_n} - P_A = P_{A_n \setminus A} \to 0$ $\mathbb{P}$-*c.s.* $\Rightarrow P_{A_n} \to P_A$ $\mathbb{P}$-*c.s.*

PROPOSICIÓN 2.3 *Sea $\mathcal{P}$ un proceso de Dirichlet en $(\Theta, \mathcal{A})$ con parámetro α y Q una medida de probabilidad en $(\Theta, \mathcal{A})$ absolutamente continua con respecto a α*

† Dados $A_1, A_2, \ldots$ eventos, si $\sum_{n=1}^{\infty} Pr(A_n) < \infty$, entonces, $Pr(\limsup A_n) = 0$, y si además los eventos son independientes, $\sum_{n=1}^{\infty} Pr(A_n) = \infty$ implica $Pr(\limsup A_n) = 1$.

$(Q \ll \alpha)$[‡]. Entonces, para cualquier entero positivo m, $A_1, A_2, \ldots, A_m \in \mathcal{A}$ y ϵ positiva

$$\mathbb{P}\{|P_{A_i} - Q(A_i)| < \epsilon, i = 1, \ldots, m\} > 0.$$

DEMOSTRACIÓN. Definiendo a $B(\nu_1, \ldots, \nu_m)$ como en (2.5) y notando que

$$\mathbb{P}[|P_{A_i} - Q(A_i)| < \epsilon, i = 1, \ldots m] \geq \mathbb{P}\left[\sum_{\{(\nu_1,\ldots,\nu_m):\nu_i=1\}} |P_{B(\nu_1,\ldots,\nu_m)} - Q(B(\nu_1, \ldots, \nu_m))| < \epsilon, i = 1, \ldots, m\right],$$

basta probar que

$$\mathbb{P}[|P_{B(\nu_1,\ldots,\nu_m)} - Q(B(\nu_1, \ldots, \nu_m))| < 2^{-m}\epsilon, \forall(\nu_1, \ldots, \nu_m)] > 0.$$

Si $\alpha(B_{\nu_1,\ldots,\nu_m}) = 0$, entonces $Q(B(\nu_1, \ldots, \nu_m)) = 0$ y $P_{B(\nu_1,\ldots,\nu_m)} = 0$ con probabilidad 1, de lo que $|P_{B(\nu_1,\ldots,\nu_m)} - Q(B_{\nu_1,\ldots,\nu_m})| = 0$ con probabilidad 1. Por lo que si $D = [(\nu_1, \ldots, \nu_m) : \alpha(B(\nu_1, \ldots, \nu_m)) > 0]$, entonces la probabilidad anterior es igual a,

$$\mathbb{P}[2^{-m}\epsilon - Q(B(\nu_1, \ldots, \nu_m)) < P_{B(\nu_1,\ldots,\nu_m)} < 2^{-m}\epsilon + Q(B(\nu_1, \ldots, \nu_m)), (\nu_1, \ldots, \nu_m) \in D] = \int_{\times\{(2^{-m}\epsilon - Q(B(\nu_1,\ldots,\nu_m)), 2^{-m}\epsilon + Q(B(\nu_1,\ldots,\nu_m))):(\nu_1,\ldots,\nu_m)\in D\}} d\mathcal{D}(\{\alpha(B(\nu_1, \ldots, \nu_m)) : (\nu_1, \ldots, \nu_m) \in D\}) > 0.$$

Por lo que la proposición queda demostrada.

◁

Recordemos que si μ es una medida, su soporte, sop μ, es el conjunto cerrado más pequeño tal que su complemento tiene medida cero, *i.e.*,

$$\text{sop } \mu \equiv \bigcap\{F : F \text{ es cerrado y } \mu(F^c) = 0\}.$$

Es fácil demostrar que x está en el soporte de μ si y sólo si para todo abierto V tal que $x \in V$ se tiene que $\mu(V) > 0$.

Considerando al espacio $([0,1]^{\mathcal{A}}, \mathcal{BF})$ con la topología de la convergencia puntual ($Q_n \to Q$, si $Q_n(A) \to Q(A)$ $\forall A \in \mathcal{A}$) tenemos que si Q es absolutamente continua

‡ La medida μ es absolutamente continua con respecto a la medida ν ($\mu \ll \nu$), si $\nu(A) = 0$ implica que $\mu(A) = 0$ para toda A en el σ-álgebra.

con respecto a α y V es un abierto en $[0,1]^{\mathcal{A}}$ tal que $Q \in V$, entonces existe un conjunto abierto básico de la topología de la convergencia puntual B contenido en V, es decir, existen $m \in \mathbb{N}$, $A_1, \ldots A_m \in \mathcal{A}$ y $\epsilon > 0$ tales que

$$B \equiv [|P_{A_i} - Q(A_i)| < \epsilon, i = 1, \ldots, m] \subset V,$$

y por la Proposición 2.3 tenemos:

$$0 < \mathbb{P}(B) \leqslant \mathbb{P}(V),$$

es decir, si $Q \ll \alpha$ entonces Q está en el soporte de $\mathbb{P}$. Notando que si $Q \not\ll \alpha$ existe A en $\mathcal{A}$ tal que $Q(A) > 0$ y $\alpha(A) = 0$, que, por la Proposición 1.1, implica que $P_A = 0$ $\mathbb{P}$-*c.s.*, tenemos que Q pertence al abierto $[|P_A - Q(A)| < \epsilon]$ con $0 < \epsilon < Q(A)$ y además $\mathbb{P}[|P_A - Q(A)| < \epsilon] = \mathbb{P}[|Q(A)| < \epsilon] = 0$, es decir, si Q no es absolutamente continua con respecto a α entonces no está en el soporte de $\mathbb{P}$.

Definición 2.4 *Sea $\mathcal{P} = \{P_A : A \in \mathcal{A}\}$ una medida de probabilidad aleatoria y $\theta_1, \theta_2, \ldots, \theta_n : ([0,1]^{\mathcal{A}}, \mathcal{BF}) \to (\Theta, \mathcal{A})$ medibles. Decimos que $\theta_1, \theta_2, \ldots, \theta_n$ es una muestra aleatoria de tamaño n de una realización de $\mathcal{P}$, si para todo $m = 1, 2, \ldots$ y $A_1, A_2, \ldots, A_m$, $C_1, \ldots C_n \in \mathcal{A}$*

$$\mathbb{P}[\theta_1 \in C_1, \ldots, \theta_n \in C_n | P_{A_1}, \ldots, P_{A_m}, P_{C_1}, \ldots, P_{C_n}] = \prod_{j=1}^{n} P_{C_j} \quad \mathbb{P} - c.s. \tag{2.11}$$

En otras palabras, $\theta_1, \theta_2, \ldots, \theta_n$ es una muestra aleatoria de tamaño n de una realización de $\mathcal{P}$, si, dado $P_{C_1}, \ldots, P_{C_n}$, los eventos $[\theta_1 \in C_1], \ldots, [\theta_n \in C_n]$ son independientes del resto del proceso e independientes entre ellos, con $\mathbb{P}[\theta_j \in C_j | P_{C_1}, \ldots, P_{C_n}] = P_{C_j}$ *c.s.* para $j = 1, \ldots, n$. Esta definición determina la distribución de $\theta_1, \ldots, \theta_n, P_{A_1}, \ldots, P_{A_m}$, una vez que el proceso está dado, ya que

$$\mathbb{P}[\theta_1 \in C_1, \ldots, \theta_n \in C_n, P_{A_1} \leqslant y_1, \ldots, P_{A_m} \leqslant y_m], \quad y_1, \ldots, y_m \in [0,1], \tag{2.12}$$

puede encontrarse integrando (2.11) con respecto a la distribución de $P_{A_1}, \ldots, P_{A_m}$, $P_{C_1}, \ldots, P_{C_n}$ sobre el conjunto $[0, y_1] \times \ldots \times [0, y_m] \times [0,1], \ldots \times [0,1]$. Las condiciones de consistencia de Kolmogorov pueden ser fácilmente verificadas para mostrar que (2.12) determina una medida de probabilidad sobre $(\Theta^n \times [0,1]^{\mathcal{A}}, \mathcal{A}^n \times \mathcal{BF})$.

PROPOSICIÓN 2.4 *Sea $\mathcal{P}$ un proceso Dirichlet en $(\Theta, \mathcal{A})$ con parámetro α y sea θ una muestra aleatoria de tamaño 1 de una realización de $\mathcal{P}$. Entonces para $A \in \mathcal{A}$,*

$$\mathbb{P}[\theta \in A] = \frac{\alpha(A)}{\alpha(\Theta)}$$

DEMOSTRACIÓN. Como $\mathbb{P}[\theta \in A|P_A] = P_A$ *c.s.*,

$$\mathbb{P}[\theta \in A] = E(1_{\theta \in A}) = E(E(1_{\theta \in A}|P_A)) = E(\mathbb{P}(\theta \in A|P_A)) = E(P_A) = \frac{\alpha(A)}{\alpha(\Theta)}.$$

◁

PROPOSICIÓN 2.5 *Sea $\mathcal{P}$ un proceso Dirichlet en $(\Theta, \mathcal{A})$ con parámetro α y sea θ una muestra aleatoria de tamaño 1 de una realización de $\mathcal{P}$. Sea $\{B_1, \ldots, B_k\}$ una partición medible de Θ y $A \in \mathcal{A}$. Entonces*

$$\mathbb{P}[\theta \in A, P_{B_1} \leqslant y_1, \ldots, P_{B_k} \leqslant y_k] = \sum_{j=1}^{k} \frac{\alpha(B_j \cap A)}{\alpha(\Theta)} D(y_1, \ldots, y_k|\alpha_1^{(j)}, \ldots, \alpha_k^{(j)}), \qquad (2.13)$$

donde $D(y_1, \ldots, y_k|\alpha_1, \ldots, \alpha_k)$ es la función de distribución Dirichlet $D(\alpha_1, \ldots, \alpha_k)$, y

$$\alpha_i^{(j)} = \begin{cases} \alpha(B_i), & si\ i \neq j \\ \alpha(B_i) + 1, & si\ i = j. \end{cases}$$

DEMOSTRACIÓN. Definamos $B_{j,1} = B_j \cap A$ y $B_{j,0} = B_j \cap A^c$ para $j = 1, \ldots, k$. Sea $Y_{j,\nu} = P_{B_{j,\nu}}$ para $j = 1, \ldots, k$ y $\nu = 0, 1$. Entonces de (2.11)

$$\begin{aligned} &\mathbb{P}[\theta \in A|Y_{j,\nu}, j = 1, \ldots, k,\ \nu = 0, 1]] \\ &= \mathbb{P}\left[\theta \in \bigcup_{j=1}^{k} B_{j,1}|Y_{j,\nu}, j = 1, \ldots, k,\ \nu = 0, 1\right] \\ &= \sum_{j=1}^{k} \mathbb{P}[\theta \in B_{j,1}|Y_{j,\nu}, j = 1, \ldots, k,\ \nu = 0, 1] \\ &= \sum_{j=1}^{k} Y_{j,1} \quad c.s. \end{aligned} \qquad (2.14)$$

Por lo tanto, para $y_{j,\nu} \in [0,1]$, $j = 1, \ldots, k,\ \nu = 0, 1$

$$\mathbb{P}[\theta \in A, Y_{j,\nu} \leqslant y_{j,\nu}, j = 1, \ldots, k,\ \nu = 0, 1]$$

$$= \int_{\times\{[0,y_{j,\nu}]:j=1,\ldots,k,\ \nu=0,1\}} \mathbb{P}[\theta \in A | Y_{1,0} = y'_{1,0}, \ldots, Y_{k,1} = y'_{k,1}] dP_{Y_{1,0},\ldots,Y_{k,1}}(y'_{1,0}, \ldots, y'_{k,1})$$

$$= \int_{\times\{[0,y_{j,\nu}]:j=1,\ldots,k,\ \nu=0,1\}} \sum_{j=1}^{k} y'_{j,1} dD(y'_{1,0}, \ldots, y'_{k,1} | \boldsymbol{\alpha}(B_{1,0}), \ldots, \boldsymbol{\alpha}(B_{k,1}))$$

$$= \sum_{j=1}^{k} \int_{\times\{[0,y_{j,\nu}]:j=1,\ldots,k,\ \nu=0,1\}} y'_{j,1} dD(y'_{1,0}, \ldots, y'_{k,1} | \boldsymbol{\alpha}(B_{1,0}), \ldots, \boldsymbol{\alpha}(B_{k,1}))$$

y de la ecuación (2.4) de la Segunda Sección

$$= \sum_{j=1}^{k} \frac{\boldsymbol{\alpha}(B_{j,1})}{\boldsymbol{\alpha}(\Theta)} D(y | \boldsymbol{\alpha}^{(j)}),$$

donde $y = (y_{1,0}, \ldots, y_{k,0}, y_{1,1}, \ldots, y_{k,1})$, $\boldsymbol{\alpha}^{(j)} = (\boldsymbol{\alpha}^{(j)}_{1,0}, \ldots, \boldsymbol{\alpha}^{(j)}_{k,0}, \boldsymbol{\alpha}^{(j)}_{1,1}, \ldots, \boldsymbol{\alpha}^{(j)}_{k,1})$ y

$$\boldsymbol{\alpha}^{(j)}_{i,1} = \begin{cases} \boldsymbol{\alpha}(B_{i,1}) + 1, & \text{si } i = j, \\ \boldsymbol{\alpha}(B_{i,1}), & \text{si } i \neq j \end{cases}$$

$$\boldsymbol{\alpha}^{(j)}_{i,0} = \boldsymbol{\alpha}(B_{i,0}).$$

Por la primera propiedad de la distribución Dirichlet, como $P_{B_j} = Y_{j,0} + Y_{j,1}$ *c.s.* y como el proceso para encontrar las distribuciones marginales es lineal se sigue el resultado.

◁

Ahora veamos cuál es la distribución del proceso Dirichlet dada una muestra $\theta_1, \ldots, \theta_n$ de tamaño n de una realización de $\mathcal{P}$. Resulta que ésta es también Dirichlet.

Teorema 2.1 *Sea $\mathcal{P}$ un proceso Dirichlet en $(\Theta, \mathcal{A})$ con parámetro $\boldsymbol{\alpha}$ y $\theta_1, \theta_2, \ldots, \theta_n$ una muestra aleatoria de tamaño n de una realización de $\mathcal{P}$. Entonces la distribución condicional de $\mathcal{P}$ dado $\theta_1, \theta_2, \ldots, \theta_n$ es un proceso Dirichlet con parámetro $\boldsymbol{\alpha} + \sum_{i=1}^{n} \delta_{\theta_i}$. En notación, si $\mathcal{P} = \mathcal{DP}(\boldsymbol{\alpha})$, entonces $\mathcal{P}|\theta_1, \ldots, \theta_n = \mathcal{DP}(\boldsymbol{\alpha} + \sum_{i=1}^{n} \delta_{\theta_i})$.*

Demostración. Basta demostrarlo para $n = 1$, ya que con esto el teorema se prueba trivialmente por inducción para cualquier n.

Sea $\{B_1, \ldots, B_k\}$ una partición medible de Θ y $A \in \mathcal{A}$, como las distribuciones marginales de una distribución condicional de un proceso son las distribuciones condicionales de las marginales †, debemos demostrar que la distribución condicional conjunta

† Es decir, si $X = \{X_t : t \in T\}$ es un proceso estocástico, Y una variable aleatoria y $Z = \{Z_t = X_t | Y : t \in T\}$, entonces, dados $k \in \mathbb{N}$, $A_1, \ldots, A_k$ en el σ-álgebra y $t_1, \ldots, t_k \in T$, $Pr[X_{t_1} \in A_1, \ldots, X_{t_k} \in A_k | Y] = Pr[Z_{t_1} \in A_1, \ldots, Z_{t_k} \in A_k]$,

de $P_{B_1}, \ldots, P_{B_k}$ dada θ, una muestra aleatoria de tamaño 1 de una realización de $\mathcal{P}$, tiene función de distribución

$$D(y_1, \ldots, y_k | \boldsymbol{\alpha}(B_1) + \delta_\theta(B_1), \ldots, \boldsymbol{\alpha}(B_k) + \delta_\theta(B_k)). \tag{2.15}$$

Esto puede ser probado mostrando que la integral de (2.15) con respecto a la distribución marginal de θ sobre A es igual a la probabilidad (2.13). Usando la distribución marginal de θ encontrada en la Proposición 2.4 calculamos

$$\begin{aligned}
&\int_A D(y_1, \ldots, y_k | \boldsymbol{\alpha}(B_1) + \delta_\theta(B_1), \ldots, \boldsymbol{\alpha}(B_k) + \delta_\theta(B_k)) d\boldsymbol{\alpha}(x)/\boldsymbol{\alpha}(\Theta) \\
&\qquad = \sum_{j=1}^{k} \int_{B_j \cap A} D(y_1, \ldots, y_k | \boldsymbol{\alpha}_1^{(j)}, \ldots, \boldsymbol{\alpha}_k^{(j)}) d\boldsymbol{\alpha}(x)/\boldsymbol{\alpha}(\Theta) \\
&\qquad = \sum_{j=1}^{k} \frac{\boldsymbol{\alpha}(B_j \cap A)}{\boldsymbol{\alpha}(\Theta)} D(y_1, \ldots, y_k | \boldsymbol{\alpha}_1^{(j)}, \ldots, \boldsymbol{\alpha}_k^{(j)}),
\end{aligned}$$

completando la demostración.

◁

La siguente proposición, que utilizaremos para demostrar que el proceso Dirichlet, cuando $(\Theta, \mathcal{A})$ es la recta real con la σ-álgebra de Borel, es una medida de probabilidad discreta $\mathbb{P}$-*c.s.*, nos indica que el proceso Dirichlet tiene cierta "memoria", es decir, la probabilidad de tener una observación repetida es positiva.

PROPOSICIÓN 2.6 *Si $\mathcal{P}$ es un proceso Dirichlet en $(\Theta, \mathcal{A})$ con parámetro $\boldsymbol{\alpha}$ y $\boldsymbol{\alpha}$ es una medida no atómica en Θ* † *y si $\theta_1, \ldots, \theta_n$ es una muestra aleatoria de tamaño n de una realización de $\mathcal{P}$ entonces,*

$$\mathbb{P}[\theta_n \neq \theta_1, \theta_n \neq \theta_2, \ldots, \theta_n \neq \theta_{n-1}] = \frac{\boldsymbol{\alpha}(\Theta)}{\boldsymbol{\alpha}(\Theta) + n - 1}. \tag{2.16}$$

DEMOSTRACIÓN. Por el Teorema 2.1 sabemos que $\mathcal{P}$ dado $\theta_1 = x_1, \ldots, \theta_{n-1} = x_{n-1}$ es un

† μ es una medida no atómica si para todo A en la σ-álgebra tal que $\mu(A) > 0$ existe $B \subset A$ tal que $\mu(B)$ y $\mu(A \setminus B)$ son positivos.

proceso Dirichlet con parámetro $\alpha + \sum_{i=1}^{n-1} \delta_{x_i}$. Utilizando la Proposición 2.4 obtenemos,

$$\begin{aligned}
&\mathbb{P}[\theta_n \neq \theta_1, \ldots, \theta_n \neq \theta_{n-1}] \\
&= \int_{[0,1]^{\mathcal{A}}} \mathbb{P}[\theta_n \neq \theta_1, \ldots, \theta_n \neq \theta_{n-1} | \theta_1, \ldots, \theta_{n-1}] \, d\mathbb{P} \\
&= \int_{\Theta^{n-1}} \mathbb{P}[\theta_n \notin \{x_1, \ldots, x_{n-1}\} | \theta_1 = x_1, \ldots, \theta_{n-1} = x_{n-1}] \, d\mathbb{P}_{\theta_1, \ldots, \theta_{n-1}}(x_1, \ldots, x_{n-1}) \\
&= \int_{\Theta^{n-1}} \frac{\left(\alpha + \sum_{i=1}^{n-1} \delta_{x_i}\right)(\{x_1, \ldots, x_{n-1}\}^c)}{\left(\alpha + \sum_{i=1}^{n-1} \delta_{x_i}\right)(\Theta)} \, d\mathbb{P}_{\theta_1, \ldots, \theta_{n-1}}(x_1, \ldots, x_{n-1}) \\
&= \int_{\Theta^{n-1}} \frac{\left(\alpha + \sum_{i=1}^{n-1} \delta_{x_i}\right)(\Theta) - \left(\alpha + \sum_{i=1}^{n-1} \delta_{x_i}\right)(\{x_1, \ldots, x_{n-1}\})}{\left(\alpha + \sum_{i=1}^{n-1} \delta_{x_i}\right)(\Theta)} \, d\mathbb{P}_{\theta_1, \ldots, \theta_{n-1}}(x_1, \ldots, x_{n-1}) \\
&= \int_{\Theta^{n-1}} \frac{(\alpha(\Theta) + n - 1) - (n-1)}{\alpha(\Theta) + n - 1} \, d\mathbb{P}_{\theta_1, \ldots, \theta_{n-1}}(x_1, \ldots, x_{n-1}) \\
&= \frac{\alpha(\Theta)}{\alpha(\Theta) + n - 1}.
\end{aligned}$$

◁

Por otro lado, dada una muestra aleatoria $\theta_1, \theta_2, \ldots, \theta_n$ de tamaño n de una realización de un proceso Dirichlet $\mathcal{P}$ con parámetro α, definimos a W_i como la variable indicadora que es 1 si obtenemos un valor de θ distinto de $\theta_1, \ldots, \theta_{i-1}$ en la i-ésima observación y cero en otro caso, y a Z_n como la cantidad de valores distintos en la muestra, es decir, $Z_n = \sum_{i=1}^{n} W_i$. Por la proposición anterior tenemos que W_i es independiente de W_j $(i \neq j)$ y que $\mathbb{P}[W_i = 1] = \alpha(\Theta)/(\alpha(\Theta) + i - 1)$. Si bien esta última probabilidad es decreciente en i, $E(Z_n) = \alpha(\Theta) \sum_{i=1}^{n} 1/(\alpha(\Theta) + i - 1) \approx \alpha(\Theta) \log[(n + \alpha(\Theta))/(\alpha(\Theta) + i - 1)]$ que tiende a infinito cuando n lo hace. De hecho Korwar y Hollander (1973) probaron que $Z_n \to \infty$ $\mathbb{P}$-$c.s.$ cuando $n \to \infty$. Por lo tanto, aunque valores distintos de θ son cada vez más raros se asegura su incremento. De hecho, si definimos los polinomios,

$$\begin{aligned}
A_1(x) &= x \\
A_2(x) &= (x+1)A_1(x) = x(x+1) \\
&\vdots \\
A_n(x) &= (x+n-1)A_{n-1}(x) = x(x+1)\ldots(x+n-1),
\end{aligned}$$

es decir, $A_n(x)$ es un polimonio de grado n en x con coeficientes enteros que escribimos:

$$A_n(x) = {}_n a_1 x + {}_n a_2 x^2 + \ldots + {}_n a_n x^n,$$

cuyos coeficientes son los valores absolutos de los números de Stirling de primera clase tabulados en Abramowitz y Stegun (1964), p. 833. Entonces la probabilidad de observar k valores distintos en una muestra de tamaño n de una realización de $\mathcal{P}$ está dada por,

$$\mathbb{P}[Z_n = k] = \frac{{}_n a_k \boldsymbol{\alpha}(\Theta)^k}{A_n(\boldsymbol{\alpha}(\Theta))}. \tag{2.17}$$

La demostración se sigue por inducción observando que,

$$\mathbb{P}[Z_1 = 1] = 1 = \frac{{}_1 a_1 \boldsymbol{\alpha}(\Theta)^1}{A_1(\boldsymbol{\alpha}(\Theta))}.$$

Suponiendo (2.17) válida para valores menores que n y $1 \leqslant k \leqslant n-1$, tenemos,

$$\begin{aligned}
\mathbb{P}[Z_n = k] &= \mathbb{P}[Z_{n-1} = k]\mathbb{P}[W_n = 0] + \mathbb{P}[Z_{n-1} = k-1]\mathbb{P}[W_n = 1] \\
&= \frac{{}_{n-1}a_k \boldsymbol{\alpha}(\Theta)^k}{A_{n-1}(\boldsymbol{\alpha}(\Theta))} \frac{n-1}{\boldsymbol{\alpha}(\Theta)+n-1} + \frac{{}_{n-1}a_{k-1}\boldsymbol{\alpha}(\Theta)^{k-1}}{A_{n-1}(\boldsymbol{\alpha}(\Theta))} \frac{\boldsymbol{\alpha}(\Theta)}{\boldsymbol{\alpha}(\Theta)+n-1} \\
&= ({}_{n-1}a_k(n-1) + {}_{n-1}a_{k-1}) \frac{\boldsymbol{\alpha}(\Theta)^k}{A_n(\boldsymbol{\alpha}(\Theta))},
\end{aligned}$$

y, observando, como fácilmente se puede verificar, que ${}_{n-1}a_k(n-1) + {}_{n-1}a_{k-1} = {}_n a_k$ se demuestra (2.17) para n y $1 \leqslant k \leqslant n-1$. Si $k = n$, la expresión (2.17) se sigue directamente notando que ${}_n a_n = 1$ y

$$\mathbb{P}[Z_n = n] = \mathbb{P}[W_i = 1, i, \ldots, n] = \prod_{i=1}^{n} \mathbb{P}[W_i = 1] = \prod_{i=1}^{n} \frac{\boldsymbol{\alpha}(\Theta)}{\boldsymbol{\alpha}(\Theta)+i-1} = \frac{{}_n a_n \boldsymbol{\alpha}(\Theta)^n}{A_n(\boldsymbol{\alpha}(\Theta))},$$

demostrando el resultado. El hecho importante del resultado anterior es que Z_n depende únicamente de la magniud de $\boldsymbol{\alpha}(\Theta)$ y no de la forma de $\boldsymbol{\alpha}$.

2.4 *Caso* $(\Theta, \mathcal{A}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$

Sea $(\Theta, \mathcal{A}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, $\boldsymbol{\alpha}$ una medida σ-aditiva en $\mathbb{R}$ y $\{P_A : A \in \mathcal{B}(\mathbb{R})\}$ el proceso Dirichlet con parámetro $\boldsymbol{\alpha}$. Si definimos $F_t : [0,1]^{\mathcal{B}(\mathbb{R})} \to [0,1]$ como

$$F_t\, f \equiv P_{(-\infty,t]}\, f = f(-\infty, t], \quad \forall f \in [0,1]^{\mathcal{B}(\mathbb{R})},$$

entonces $F \equiv \{F_t : t \in \mathbb{R}\}$ es un proceso estocástico con valores en $[0,1]$ y conjunto de índices $\mathbb{R}$.

Demostraremos que existe $F^* = \{F_t^* : t \in \mathbb{R}\}$ modificación de $F = \{F_t : t \in \mathbb{R}\}$ tal que F^* es función de distribución con probabilidad 1. (F^* es modificación de F si $F_t^* = F_t$ con

probabilidad 1 para todo t en $\mathbb{R}$; en particular F^* es equivalente a F, es decir tienen las mismas distribuciones finito dimensionales).

Para toda t en $\mathbb{R}$, como $(-\infty, t+1/n] \searrow (-\infty, t]$, por (2.10) tenemos

$$F_{t+1/n} \searrow F_t \quad c.s. \tag{2.18}$$

Entonces

$$C \equiv \left[\lim_{n\to\infty} F_{t+1/n} = F_t, \forall t \in \mathbb{Q}\right] \qquad \text{tiene } \mathbb{P}\text{-medida 1.}$$

Por otro lado si $s < t$, $(-\infty, s] \subset (-\infty, t]$ de lo que tenemos que $F_s \leqslant F_t$ $\mathbb{P} - c.s.$ y de ahí,

$$D \equiv [F_s \leqslant F_t \ \forall s, t \in \mathbb{Q}, s < t]$$

tiene $\mathbb{P}$-medida 1. Si $s_n \searrow s$ con $s \in \mathbb{Q}$ y $\{s_n\} \subset \mathbb{Q}$, entonces toda subsucesión de $\{s_n\}$, $\{s_{n_k}\}$ contiene otra subsucesión $\{s_{n_{k_j}}\}$ tal que $s \leqslant s_{n_{k_j}} \leqslant s + 1/n$, y como

$$[F_s \leqslant F_{s_{n_{k_j}}} \leqslant F_{s+1/n}] \subset C \cap D,$$

entonces $F_{s_{n_{k_j}}} \searrow F_s$ de lo que concluimos que $F_{s_n} \searrow F_s$, es decir

$$E \equiv [F_{s_n} \searrow F_s, \ \forall \{s_n\} \subset \mathbb{Q}, s \in \mathbb{Q} \text{ tal que } s_n \searrow s]$$

tiene $\mathbb{P}$-medida 1. Definamos

$$F_t^* \equiv \begin{cases} F_t, & \text{si } \ t \in \mathbb{Q} \\ \inf\{F_s | s > t, s \in \mathbb{Q}\}, & \text{si } \ t \notin \mathbb{Q}. \end{cases} \tag{2.19}$$

Por construcción las trayectorias de F^* son no decrecientes en E. Además si $s_n \searrow s$, existe $\{s_n'\} \subset \mathbb{Q}$ tal que $s_n' \searrow s$ y $s \leqslant s_n \leqslant s_n'$ lo que implica que $F_s^* \leqslant F_{s_n}^* \leqslant F_{s_n'}^*$ en E, por lo que $F_{s_n}^* \searrow F_s^*$ en E, es decir, F^* tiene trayectorias continuas por la derecha y no decrecientes $\mathbb{P}$-$c.s.$, y de esto se sigue que tiene límites por la izquierda.

Por otro lado $-n \searrow -\infty$, así que por la Proposición 2.2 y como $(-\infty, -n] \searrow \emptyset$, $F_{-n}^* = F_{-n} \searrow 0$ y por tener F^* trayectorias no decrecientes $\mathbb{P} - c.s.$ tenemos

$$\lim_{t\to-\infty} F_t^* = 0 \quad \mathbb{P} - c.s.$$

como $n \nearrow \infty$, $(-\infty, n] \nearrow \mathbb{R}$, y como $P(\mathbb{R}) = 1$ $\mathbb{P} - c.s.$, por (2.10), se sigue que $F_n^* = F_n \nearrow 1$ y por tener F^* trayectorias no decrecientes:

$$\lim_{t\to\infty} F_t^* = 1 \quad \mathbb{P} - c.s.$$

Con todo lo anterior de esta sección hemos probado que las trayectorias de F^* son funciones de distribución, y de (2.18) y (2.19), tenemos que $F_t^* = F_t$ con probabilidad 1 para toda t en $\mathbb{R}$, es decir, F^* es una modificación de F, con lo que probamos que $\{P_A : A \in \mathcal{B}(\mathbb{R})\}$ es medida de probabilidad $\mathbb{P} - c.s.$, o equivalentemente, $\mathbb{P} : \mathcal{BF} \to [0,1]$ es una medida de probabilidad en $[0,1]^{\mathcal{B}(\mathbb{R})}$, que asigna probabilidad 1 al conjunto de las medidas de probabilidad en $\mathbb{R}$. Debido a esto, en adelante, podremos pensar a $\mathbb{P}$ como una medida de probabilidad en Π, donde

$$\Pi \equiv \{p : \mathcal{B}(\mathbb{R}) \to [0,1] : p \text{ es medida de probabilidad en } \mathbb{R}\},$$

con su σ-álgebra correspondiente, $\{A \cap \Pi : A \in \mathcal{BF}\}$.

Ferguson (1973) y Blackwell y MacQueen (1973) probaron que existe un conjunto de medidas de probabilidad discretas, $\Pi_0 \subset \Pi$, tal que $\mathbb{P}$ asigna probabilidad 1 a Π_0. Sethuraman (1994) demuestra un teorema alternativo, que no solamente muestra que el proceso Dirichlet es una medida de probabilidad discreta $\mathbb{P}$-*c.s.*, sino que además da un algoritmo para simular aproximadamente del proceso Dirichlet. Lo que probaremos es un teorema de Kraster y Pratt (1986), extraido de Schervish (1995) p. 56, que implica que el proceso Dirichlet asigna probabilidad 1 al conjunto de medidas de probabilidad discretas.

Teorema 2.2 *Sea $\mathcal{P}$ el proceso Dirichlet en $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ con parámetro α no atómico, $\theta_1, \ldots, \theta_n$ una muestra aleatoria de tamaño n de una realización de $\mathcal{P}$ y definamos*

$$a_n = \mathbb{P}[\theta_n \textit{ sea distinta de } \theta_1, \theta_2, \ldots, \theta_{n-1}].$$

Si el $\lim_{n\to\infty} a_n = 0$, entonces $\mathcal{P}$ es una medida de probabilidad discreta $\mathbb{P}$-c.s.

Demostración. Definamos a $B_\epsilon = \{p \in \Pi : \exists A_p \in \mathcal{B}(\mathbb{R}) \text{ tal que } p(A_p) > \epsilon \text{ y } p(\{x\}) = 0 \ \forall x \in A_p\}$. Es fácil notar que $B_\epsilon \subset B_\delta$, si $\epsilon > \delta$, y que $B_\epsilon \nearrow B$ cuando $\epsilon \to 0$, donde

$$B = \{p \in \Pi : \exists A_p \in \mathcal{B}(\mathbb{R}) \text{ tal que } p(A_p) > 0 \text{ y } p(\{x\}) = 0 \ \forall x \in A_p\},$$

por lo que

$$B^c = \{p \in \Pi : \forall A \in \mathcal{B}(\mathbb{R}) \text{ tal que } p(A) > 0, \exists x \in A \text{ tal que } p(\{x\}) > 0\},$$

es decir,

$$B^c = \{p \in \Pi : p \text{ es una medida de probabilidad discreta en } (\mathbb{R}, \mathcal{B}(\mathbb{R}))\}.$$

Para demostrar que $\mathbb{P}(B^c) = 1$, basta con demostrar que $\mathbb{P}(B) = 0$, y esto se hace probando que $\mathbb{P}(B_\epsilon) = 0$ para toda ϵ positiva . Definamos

$$A = \{(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n : x_n \neq x_1, x_n \neq x_2, \ldots, x_n \neq x_{n-1}\},$$

entonces,

$$\begin{aligned}
a_n &= \mathbb{P}[\theta_n \text{ sea distinta de } \theta_1, \theta_2, \ldots, \theta_{n-1}] \\
&= \mathbb{P}[(\theta_1, \ldots, \theta_n) \in A] \\
&= \int_{\Pi} \mathbb{P}[(\theta_1, \ldots, \theta_n) \in A | \theta_1, \ldots, \theta_{n-1}, \mathcal{P}]\, d\mathbb{P} \\
&= \int_{\mathbb{R}^{n-1} \times \Pi} \mathbb{P}[(\theta_1, \ldots, \theta_n) \in A | \theta_1 = x_1, \ldots, \theta_{n-1} = x_{n-1}, \mathcal{P} = p] \\
&\qquad\qquad d\mathbb{P}_{\theta_1, \ldots, \theta_{n-1}, \mathcal{P}}(x_1, \ldots, x_{n-1}, p) \\
&= \int_{\Pi} \int_{\mathbb{R}^{n-1}} \mathbb{P}[\theta_n \notin \{x_1, \ldots, x_{n-1}\} | \mathcal{P} = p]\, d\mathbb{P}_{\theta_1, \ldots, \theta_{n-1}}(x_1, \ldots, x_{n-1})\, d\mathbb{P}_{\mathcal{P}}(p) \\
&\geqslant \int_{B_\epsilon} \int_{\mathbb{R}^{n-1}} P[\theta_n \notin \{x_1, \ldots, x_{n-1}\}]\, d\mathbb{P}_{\theta_1, \ldots, \theta_{n-1}}(x_1, \ldots, x_{n-1})\, d\mathbb{P}_{\mathcal{P}}(p).
\end{aligned}$$

Notemos que si $p \in B_\epsilon$,

$$\begin{aligned}
&\int_{\mathbb{R}^{n-1}} p[\theta_n \notin \{x_1, \ldots, x_{n-1}\}]\, d\mathbb{P}_{\theta_1, \ldots, \theta_{n-1}}(x_1, \ldots, x_{n-1}) \\
&\qquad \geqslant \int_{A_p} p[\theta_n \notin \{x_1, \ldots, x_{n-1}\}]\, d\mathbb{P}_{\theta_1, \ldots, \theta_{n-1}}(x_1, \ldots, x_{n-1}) \\
&\qquad = \int_{A_p} 1\, d\mathbb{P}_{\theta_1, \ldots, \theta_{n-1}}(x_1, \ldots, x_{n-1}) = p(A_p) > \epsilon.
\end{aligned}$$

Por lo que

$$a_n \geqslant \int_{B_\epsilon} \epsilon\, d\mathbb{P}_{\mathcal{P}}(p) = \epsilon \mathbb{P}(B_\epsilon).$$

Entonces si $\lim_{n\to\infty} a_n = 0$, $\mathbb{P}(B_\epsilon) = 0$ para toda ϵ positiva, con lo que se demuestra el teorema[†].

◁

COROLARIO 2.1 *El proceso Dirichlet en* $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ *con parámetro* α *no atómico es discreto* $\mathbb{P}$*-c.s.*

[†] Notemos que no se utilizó el hecho de que $\mathcal{P}$ sea un proceso Dirichlet; para que el resultado sea válido basta que $\mathcal{P}$ sea una medida de probabilidad *c.s.*

DEMOSTRACIÓN. a_n, definido como en el Teorema 2.2 es, por la Proposición 2.6, $\alpha(\Theta)/(\alpha(\Theta)+n-1)$ que tiende a cero cuando n tiende a infinito y aplicando este último teorema se obtiene el resultado.

◁

A pesar de que el proceso Dirichlet es discreto $\mathbb{P}$-*c.s.*, es decir, no cumple con la propiedad 2 dada en la Sección 2.1, en las situaciones en las que queremos una distribución inicial en la clase de las distribuciones continuas la Proposición 2.3 asegura que si Q es una medida de probabilidad absolutamente continua con respecto al parámetro del proceso α, entonces Q está en el soporte de $\mathbb{P}$, es decir, existen realizaciones del proceso (medidas de probabilidad en $(\Theta, \mathcal{B}(\mathbb{R}))$ tan cerca como se quiera a Q (considerando la topología de la convergencia puntual en Π).

2.5 *Mezclas de procesos Dirichlet*

Como ya habíamos mencionado en la Sección 2.1, el proceso Dirichlet satisface la primera y la tercera propiedades que desearíamos que satisfaciera una familia de distribuciones inicial $\mathcal{D}$ sobre el espacio de distribuciones $\mathcal{F}$ en $(\Theta, \mathcal{A})$, pero es ligeramente deficiente en la segunda. Además, hay ciertos análisis de problemas paramétricos bayesianos en los cuales al usar el proceso Dirichlet la propiedad 1(c) no se satisface. Por ejemplo, la forma de muestreo en los problemas de mezclas de distribuciones y en problemas de bioensayo puede ser de tal manera que la distribución final no es un simple proceso de Dirichlet, sin embargo ésta puede ser representada como una mezcla de procesos Dirichlet, que es, por decirlo de alguna manera, un proceso Dirichlet donde el parámetro α es aleatorio.

En esta sección revisaremos el trabajo de Antoniak (1974), definiendo lo que es la mezcla de procesos Dirichlet, derivando sus propiedades básicas y demostrando que ésta satisface la propiedad de cerradura 1(c).

Para definir la mezcla de procesos Dirichlet, necesitamos primero definir una generalización del concepto de probabilidad de transición.

DEFINICIÓN 2.5 *Sean $(\Theta, \mathcal{A})$ y $(\mathcal{U}, \mathcal{B})$ dos espacios de medida. Una medida de transición en $\mathcal{U} \times \mathcal{A}$ es una función de $\mathcal{U} \times \mathcal{A}$ al $[0, \infty)$ tal que:*

(a) Para todo $u \in \mathcal{U}$, $\alpha(u, \cdot)$ es una medida finita no nula en $(\Theta, \mathcal{A})$; y

(b) Para todo $A \in \mathcal{A}$, $\alpha(\cdot, A)$ es medible en $(\mathcal{U}, \mathcal{B})$.

Notemos que esta definición difiere de la de probabilidad de transición en que $\alpha(\cdot, \Theta)$ no es, necesariamente, idénticamente 1. Hacemos este cambio porque buscamos que $\alpha(u, \cdot)$ sea un parámetro de un proceso Dirichlet.

Definición 2.6 *Sean $(\Theta, \mathcal{A})$ un espacio de medida y $(\mathcal{U}, \mathcal{B}, H)$ uno de probabilidad, que llamaremos espacio índice y α una medida de transición en $\mathcal{U} \times \mathcal{A}$. Decimos que $\mathcal{P} = \{P_A : A \in \mathcal{A}\}$ es una mezcla de procesos Dirichlet en $(\Theta, \mathcal{A})$ con distribución de mezcla H en $(\mathcal{U}, \mathcal{B})$ y parámetro α, si para toda $k \in \mathbb{N}$, $y_1, \ldots, y_k \in [0,1]$ y $B_1, \ldots, B_k$ partición medible de Θ, tenemos*

$$\mathbb{P}[P_{B_1} \leqslant y_1, \ldots, P_{B_k} \leqslant y_k] = \int_{\mathcal{U}} D(y_1, \ldots, y_k | \alpha(u, B_1), \ldots, \alpha(u, B_k))\, dH(u),$$

y utilizaremos la notación

$$(P_{B_1}, \ldots, P_{B_k}) \sim \int_{\mathcal{U}} D(\alpha(u, B_1), \ldots, \alpha(u, B_k))\, dH(u),$$

o simplemente,

$$\mathcal{P} = \int_{\mathcal{U}} \mathcal{DP}(\alpha(u, \cdot))\, dH(u).$$

Podemos considerar al índice u como una variable aleatoria con distribución H y condicionado a u, $\mathcal{P}$ es un proceso Dirichlet con parámetro $\alpha(u, \cdot)$. De hecho podemos definir $\mathcal{U}$ como una variable aleatoria identidad y utilizaremos la notación $|u$ para denotar "dado $\mathcal{U} = u$". En notación alternativa, $u \sim H$ y $\mathcal{P}|u = \mathcal{DP}(\alpha_u)$, donde $\alpha_u = \alpha(u, \cdot)$.

Un ejemplo importante de una mezcla de procesos Dirichlet derivada de un proceso Dirichlet es el siguiente

Ejemplo 1.1. Sea $\mathcal{P}$ un proceso Dirichlet en $(\Theta, \mathcal{A})$ con parámetro α. Definamos $\alpha(u, A) = \alpha(A) + \delta_u(A)$, para $A \in \mathcal{A}$, y H una medida de probabilidad en $(\Theta, \mathcal{A})$. Entonces el proceso $\mathcal{P}^*$ que elige u de acuerdo a H y a $\mathcal{P}$ como un proceso Dirichlet con parámetro $\alpha(u, \cdot)$ es una mezcla de procesos Dirichlet. Además para $y_1, \ldots, y_k \in [0,1]$ y cualquier partición

medible $\{B_1, \ldots, B_k\}$ de Θ,

$$\begin{aligned}
&\mathbb{P}[P^*_{B_1} \leqslant y_1, \ldots, P^*_{B_k} \leqslant y_k] \\
&= \int_\Theta D(y_1, \ldots, y_k | \boldsymbol{\alpha}(u, B_1), \ldots, \boldsymbol{\alpha}(u, B_k))\, dH(u) \\
&= \sum_{i=1}^{k} \int_{B_i} D(y_1, \ldots, y_k | \boldsymbol{\alpha}(B_1) + \delta_u(B_1), \ldots, \boldsymbol{\alpha}(B_k) + \delta_u(B_k))\, dH(u) \qquad (2.20) \\
&= \sum_{i=1}^{k} H(B_i) D(y_1, \ldots, y_k | \boldsymbol{\alpha}(B_1), \ldots, \boldsymbol{\alpha}(B_i) + 1, \ldots, \boldsymbol{\alpha}(B_k)),
\end{aligned}$$

es decir,

$$(P^*_{B_1}, \ldots, P^*_{B_k}) \sim \sum_{i=1}^{k} H(B_i) \mathcal{D}(\boldsymbol{\alpha}(B_1), \ldots, \boldsymbol{\alpha}(B_i) + 1, \ldots, \boldsymbol{\alpha}(B_k)). \qquad (2.21)$$

De hecho esta relación caracteriza al tipo de mezclas dadas en este ejemplo y esto será utilizado más adelante.

Ahora definamos una muestra aleatoria de una realización de una mezcla de procesos Dirichlet que es una extensión de una del proceso Dirichlet. Es decir, si la mezcla es un simple proceso Dirichlet, entonces la definición de una muestra aleatoria de una realización de la mezcla de procesos Dirichlet coincide con la Definición 2.4.

Definición 2.7 *Sea $\mathcal{P}$ una mezcla de procesos Dirichlet en $(\Theta, \mathcal{A})$ con distribución de mezcla H en el espacio índice $(\mathcal{U}, \mathcal{B})$ y medida de transición $\boldsymbol{\alpha}$ en $\mathcal{U} \times \mathcal{A}$. Decimos que $\theta_1, \ldots, \theta_n$ es una muestra aleatoria de tamaño n de una realización de $\mathcal{P}$ si para cualquier $m = 1, 2, \ldots$ y $A_1, \ldots, A_m, C_1, \ldots, C_n \in \mathcal{A}$ tenemos*

$$\mathbb{P}[\theta_1 \in C_1, \ldots, \theta_n \in C_n | u, P_{A_1}, \ldots, P_{A_m}, P_{C_1}, \ldots, P_{C_n}] = \prod_{i=1}^{n} P_{C_i}. \qquad c.s., \qquad (2.22)$$

Es decir, estamos pensando que $\theta_1, \ldots, \theta_n$ es una muestra aletoria de una distribución G generada por un proceso Dirichlet con parámetro $\boldsymbol{\alpha}(u, \cdot)$, y lo denotaremos por $\theta_1, \ldots, \theta_n \sim G$ y $G|u \sim \mathcal{DP}(\boldsymbol{\alpha}_u)$.

La definición anterior determina la distribución conjunta de $\theta_1, \ldots, \theta_n, P_{A_1}, \ldots, P_{A_m}$ ya que,

$$\mathbb{P}[\theta_1 \in C_1, \ldots, \theta_n \in C_n, P_{A_1} \leqslant y_1, , \ldots, P_{A_m} \leqslant y_m] \qquad (2.23)$$

puede ser encontrada integrando (2.22) con respecto a la distribución condicional conjunta de $P_{A_1}, \ldots, P_{A_m}, P_{C_1}, \ldots, P_{C_n}$ dado u sobre el conjunto $[0, y_1] \times \ldots \times [0, y_m] \times [0,1] \times \ldots \times [0,1]$ y posteriormente la función de u resultante sobre $\mathcal{U}$ con respecto a H. Una consecuencia inmediata de esta definición es

PROPOSICIÓN 2.7 *Si* $\mathcal{P} = \int_{\mathcal{U}} D(\boldsymbol{\alpha}(u, \cdot))\, dH(u)$ *y* θ *es una m.a. de tamaño 1 de una realización de* $\mathcal{P}$*, entonces para todo* $A \in \mathcal{A}$*,*

$$\mathbb{P}[\theta \in A] = \int_{\mathcal{U}} \frac{\boldsymbol{\alpha}(u, A)}{\boldsymbol{\alpha}(u, \Theta)}\, dH(u). \tag{2.24}$$

DEMOSTRACIÓN. Por la Proposición 2.4,

$$\mathbb{P}[\theta \in A | u] = \frac{\boldsymbol{\alpha}(u, A)}{\boldsymbol{\alpha}(u, \Theta)},$$

e integrando sobre el conjunto $\mathcal{U}$ con respecto a H obtenemos el resultado.

◁

TEOREMA 2.3 *Sea* $\mathcal{P}$ *un proceso Dirichlet en* $(\Theta, \mathcal{A})$ *con parámetro* $\boldsymbol{\alpha}$*,* θ *una muestra aleatoria de tamaño* 1 *de una realización de* $\mathcal{P}$ *y* $A \in \mathcal{A}$ *tal que* $\boldsymbol{\alpha}(A) > 0$*. Entonces la distribución condicional de* $\mathcal{P}$ *dado* $\theta \in A$ *es una mezcla de procesos Dirichlet en* $(\Theta, \mathcal{A})$ *con espacio índice* $(A, A \cap \mathcal{A})$*, medida de transición* $\boldsymbol{\alpha}$ *en* $A \times (A \cap \mathcal{A})$ *y distribución de mezcla* H_A *dadas por* $\boldsymbol{\alpha}(u, \cdot) = \boldsymbol{\alpha} + \delta_u$ *para todo* $u \in A$ *y* $H_A(\cdot) = \boldsymbol{\alpha}(\cdot)/\boldsymbol{\alpha}(A)$*.*

DEMOSTRACIÓN. Sean $y_1, \ldots, y_k \in [0,1]$ y $\{B_1, \ldots, B_k\}$ una partición medible de Θ, entonces, utilizando las Proposiciones 2.4 y 2.5 tenemos

$$\begin{aligned}
&\mathbb{P}[P_{B_1} \leqslant y_1, \ldots, P_{B_k} \leqslant y_k | \theta \in A] \\
&= \frac{\mathbb{P}[P_{B_1} \leqslant y_1, \ldots, P_{B_k} \leqslant y_k, \theta \in A]}{\mathbb{P}[\theta \in A]} \\
&= \frac{\sum_{i=1}^{k} \{\boldsymbol{\alpha}(B_i \cap A)/\boldsymbol{\alpha}(\Theta)\} D(y_1, \ldots, y_k | \boldsymbol{\alpha}(B_1), \ldots, \boldsymbol{\alpha}(B_i) + 1, \ldots, \boldsymbol{\alpha}(B_k))}{\boldsymbol{\alpha}(A)/\boldsymbol{\alpha}(\Theta)} \\
&= \sum_{i=1}^{k} \frac{\boldsymbol{\alpha}(B_i \cap A)}{\boldsymbol{\alpha}(A)} D(y_1, \ldots, y_k | \boldsymbol{\alpha}(B_1), \ldots, \boldsymbol{\alpha}(B_i) + 1, \ldots, \boldsymbol{\alpha}(B_k)),
\end{aligned}$$

que es justamente la relación (2.21) con $H(\cdot) = \boldsymbol{\alpha}(\cdot)/\boldsymbol{\alpha}(A)$, por lo que la proposición queda demostrada.

◁

El siguiente corolario nos permite reducir algunas mezclas de procesos Dirichlet a un simple proceso Dirichlet.

COROLARIO 2.2 *Sea $\mathcal{P}$ una mezcla de procesos Dirichlet en $(\Theta, \mathcal{A})$ con espacio índice también dado por $(\Theta, \mathcal{A})$ y medida de transición $\alpha_u = \alpha + \delta_u$. Si la distribución H en el espacio índice $(\Theta, \mathcal{A})$ está dada por $H(\cdot) = \alpha(\cdot)/\alpha(\Theta)$, entonces $\mathcal{P}$ es, de hecho, un proceso Dirichlet simple en $(\Theta, \mathcal{A})$ con parámetro α. Es decir,*

$$\int_{\Theta} \mathcal{DP}(\alpha + \delta_u) \frac{\alpha(du)}{\alpha(\Theta)} = \mathcal{DP}(\alpha).$$

DEMOSTRACIÓN. Sean $y_1, \ldots, y_k \in [0,1]$ y $\{B_1, \ldots, B_k\}$ una partición medible de Θ, entonces, por el Ejemplo 1.1,

$$\mathbb{P}[P_{B_1} \leqslant y_1, \ldots, P_{B_k} \leqslant y_k] = \sum_{i=1}^{k} \frac{\alpha(B_i)}{\alpha(\Theta)} D(y_1, \ldots, y_k | \alpha(B_1), \ldots, \alpha(B_i) + 1, \ldots, \alpha(B_k)),$$

que por la Proposición 2.5 (haciendo $A = \Theta$) caracteriza al proceso Dirichlet con parámetro α.

◁

Notemos que si en el Teorema 2.3 hiciéramos $A = \Theta$, obtendríamos que $H(\cdot) = \alpha(\cdot)/\alpha(\Theta)$, pero $\theta \in \Theta$ no da ninguna información, por lo que la distribución final de $\mathcal{P}$ es la misma que la inicial, un proceso Dirichlet simple.

TEOREMA 2.4 *Sea $\mathcal{P}$ un proceso Dirichlet en $(\Theta, \mathcal{A})$ con parámetro α, θ una muestra aleatoria de tamaño 1 de una realización de $\mathcal{P}$ y $A \in \mathcal{A}$ tal que $\alpha(A) > 0$. Entonces la distribución condicional de $\mathcal{P}$ dado P_A y $\theta \in A$ es la misma que la de $\mathcal{P}$ dado P_A.*

DEMOSTRACIÓN. Dados $y_1, \ldots, y_m \in [0,1]$ y $A, A_1, \ldots, A_m \in \mathcal{A}$, por definición de muestra aleatoria de una realización de un proceso Dirichlet (Definición 2.4), tenemos

$$\mathbb{P}[\theta \in A | P_{A_1}, \ldots, P_{A_m}, P_A] = P_A, \quad \mathbb{P} - \text{c.s.}$$

Por lo que,

$$\begin{aligned}
&\mathbb{P}[P_{A_1} \leqslant y_1, \ldots, P_{A_m} \leqslant y_m | P_A, \theta \in A] \\
&= \frac{\mathbb{P}[\theta \in A, P_{A_1} \leqslant y_1, \ldots, P_{A_m} \leqslant y_m | P_A]}{\mathbb{P}[\theta \in A | P_A]} \\
&= \frac{\mathbb{P}[\theta \in A | P_{A_1} \leqslant y_1, \ldots, P_{A_m} \leqslant y_m, P_A] \mathbb{P}[P_{A_1} \leqslant y_1, \ldots, P_{A_m} \leqslant y_m | P_A]}{\mathbb{P}[\theta \in A | P_A]} \\
&= \frac{P_A}{P_A} \mathbb{P}[P_{A_1} \leqslant y_1, \ldots, PA_m \leqslant y_m | P_A] \\
&= \mathbb{P}[P_{A_1} \leqslant y_1, \ldots, P_{A_m} \leqslant y_m | P_A].
\end{aligned}$$

Por lo tanto, las distribuciones finito dimensionales de $\mathcal{P}|P_A$ y las de $\mathcal{P}|P_A, \theta \in A$ coinciden y el teorema queda demostrado.

◁

El teorema anterior nos indica que si P_A es conocido, el evento $\theta \in A$ no nos dice nada del proceso. Esto es consistente con la definición de muestra aleatoria, y como $P_\Theta = 1$ *c.s.*, confirmamos la interpretación dada en el Corolario 2.2.

Para $\mathcal{P}$, una mezcla de procesos Dirichlet en $(\Theta, \mathcal{A})$ con espacio índice $(\mathcal{U}, \mathcal{B}, H)$ y medida de transición α en $\mathcal{U} \times \mathcal{A}$, y $\theta_1, \ldots, \theta_n$, una muestra aleatoria de tamaño n de una realización de $\mathcal{P}$, necesitaremos obtener la distribución condicional de u dado $\theta_1, \ldots, \theta_n$, a partir de la ya conocida de $\mathcal{P}$ dados $u, \theta_1, \ldots, \theta_n$. Para asegurar la existencia de la primera requerimos que $(\Theta, \mathcal{A})$ y $(\mathcal{U}, \mathcal{B})$ sean espacios de Borel estándares, definidos de la siguiente manera.

Definición 2.8 *Un espacio de Borel estándar es un espacio de medida $(\Theta, \mathcal{A})$ donde $\mathcal{A}$ es numerablemente generado y para el cual existe una función bimedible (medible y con inversa medible) de Θ sobre un espacio métrico, separable y completo.*

El siguiente teorema establece que si tomamos una muestra de una realización de una mezcla de procesos Dirichlet, y si ésta es distorsionada por un error aleatorio, la distribución final del proceso es nuevamente una mezcla de procesos Dirichlet. Como veremos en las aplicaciones del tercer capítulo, este hecho ocurre frecuentemente.

Teorema 2.5 *Sea $\mathcal{P}$ una mezcla de procesos Dirichlet en un espacio de Borel estándar $(\Theta, \mathcal{A})$ con espacio índice $(\mathcal{U}, \mathcal{B})$, distribución H en $(\mathcal{U}, \mathcal{B})$ y medida de transición α en $\mathcal{U} \times \mathcal{A}$. Sea $(\mathcal{Y}, \mathcal{C})$ un espacio Borel estándar y F una probabilidad de transición*

en $\mathcal{C} \times \Theta$ al intervalo $[0,1]$. Si θ es una m.a. de tamaño 1 de una realización de $\mathcal{P}$, es decir, $\theta \sim G$, $G|u \sim \mathcal{DP}(\alpha_u)$, y además, $Y|\theta \sim F(\cdot;\theta,)$, entonces la distribución de $\mathcal{P}$ dada θ es una mezcla de procesos Dirichlet en $(\Theta, \mathcal{A})$, con espacio índice $(\Theta \times \mathcal{U}, \mathcal{A} \times \mathcal{B})$, medida de transición $\alpha_u + \delta_\theta$ en $(\Theta \times \mathcal{U}) \times \mathcal{A}$ y distribución de mezcla H_y en el espacio índice $(\Theta \times \mathcal{U}, \mathcal{A} \times \mathcal{B})$ donde H_y es la distribución condicional de (θ, u) dado $Y = y$, esto es, si

$$u \sim H, \qquad \mathcal{P}|u \sim \mathcal{DP}(\alpha_u), \qquad \mathcal{P} = \int_{\mathcal{U}} \mathcal{DP}(\alpha_u)\, dH(u),$$

$$\theta \sim G,\ G|u \sim \mathcal{DP}(\alpha_u) \qquad y \qquad Y|\theta \sim F(\cdot;\theta)$$

entonces

$$(\mathcal{P}|Y = y) = \int_{\Theta \times \mathcal{U}} \mathcal{D}(\alpha_u + \delta_\theta)\ dH_y(\theta, u).$$

Demostración. Como $\mathcal{P}$ dado (θ, u) es un proceso Dirichlet con parámetro $\alpha_u + \delta_\theta$, para cualquier partición medible de Θ $\{B_1, \ldots, B_k\}$ y para $y_1, \ldots, y_k \in [0,1]$ tenemos

$$\begin{aligned}
&\mathbb{P}[P_{B_1} \leqslant y_1, \ldots, P_{B_k} \leqslant y_k | Y = y] \\
&= \int_{\Theta \times \mathcal{U}} \mathbb{P}[P_{B_1} \leqslant y_1, \ldots, P_{B_k} \leqslant y_k | \theta, \mathcal{U} = u, Y = y]\, d\mathbb{P}_{\theta,\mathcal{U}|Y=y}(\theta, u) \\
&= \int_{\Theta \times \mathcal{U}} D(y_1, \ldots, y_k | \alpha_u + \delta_\theta(B_1), \ldots, \alpha_u + \delta_\theta(B_k))\, dH_y(\theta, u),
\end{aligned}$$

con lo que se concluye la demostración del teorema.

◁

Ilustremos este teorema con un ejemplo.

Ejemplo 1.2. Sea Y una variable aleatoria Bernoulli con $Pr[Y = 1|\theta] = \theta$, donde θ tiene como distribución inicial una mezcla de distribuciones beta: $g(\theta) = \frac{1}{2}\mathcal{B}e(\theta|1,2) + \frac{1}{2}\mathcal{B}e(\theta|2,2)$. Denotaremos, por $\mathcal{B}e[A|a,b]$ a $\int_A \mathcal{B}e(x|a,b)\, dx$.

Supongamos el modelo: $(\Theta, \mathcal{A}) = ([0,1], \mathcal{B}[0,1])$, $\mathcal{U} = \{1,2\}$ con $H(\{1\}) = H(\{2\}) = 1/2$, $\mathcal{Y} = \{0,1\}$, $\alpha(u, A) = \mathcal{B}e[A|u,2]$ para $u \in \mathcal{U}$ y $\mathcal{P}$ una mezcla de procesos Dirichlet con estos parámetros. Si calculamos $H_y(\theta, u)$ obtenemos,

$$dH_y(\theta, u) = \mathbb{P}[\theta, \mathcal{U} = u | Y = y] \propto \frac{(u+2)\Gamma(u)}{\Gamma(3-y)\Gamma(u+y)} \mathcal{B}e(\theta|u+y, 3-y)\,.$$

Ahora, si $y = 1$,

$$dH_1(\theta, 1) = \frac{3}{5}\mathcal{B}e(\theta|2, 2)$$
$$dH_1(\theta, 2) = \frac{2}{5}\mathcal{B}e(\theta|3, 2),$$

por lo que, $dH_1(\theta) = \frac{3}{5}\mathcal{B}e(\theta|2, 2) + \frac{2}{5}\mathcal{B}e(\theta|3, 2)$. De lo anterior, utilizando el Teorema 2.4, obtenemos que para $y_1, \ldots, y_k \in [0, 1]$ y $\{B_1, \ldots, B_k\}$ partición medible de $[0, 1]$,

$$\begin{aligned}
&\mathbb{P}[P_{B_1} \leqslant y_1, \ldots, P_{B_k} \leqslant y_k | Y = 1] \\
&= \int_{\Theta \times \mathcal{U}} D(y_1, \ldots, y_k | \alpha_u + \delta_\theta(B_1), \ldots, \alpha_u + \delta_\theta(B_k))\, dH_1(\theta, u) \\
&= \frac{3}{5} \int_{[0,1]} D(y_1, \ldots, y_k | \alpha_u + \delta_\theta(B_1), \ldots, \alpha_u + \delta_\theta(B_k) \mathcal{B}e(\theta|2, 2))\, d\theta \\
&\quad + \frac{2}{5} \int_{[0,1]} D(y_1, \ldots, y_k | \alpha_u + \delta_\theta(B_1), \ldots, \alpha_u + \delta_\theta(B_k)) \mathcal{B}e(\theta|3, 2))\, d\theta \\
&= \frac{3}{5} \sum_{i=1}^{k} D(y_1, \ldots, y_k | \mathcal{B}e[B_1|1, 2], \ldots, \mathcal{B}e[B_i|1, 2] + 1, \ldots, \mathcal{B}e[B_k|1, 2]) \mathcal{B}e[B_i|2, 2] \\
&\quad + \frac{2}{5} \sum_{i=1}^{k} D(y_1, \ldots, y_k | \mathcal{B}e[B_1|2, 2], \ldots, \mathcal{B}e[B_i|2, 2] + 1, \ldots, \mathcal{B}e[B_k|2, 2]) \mathcal{B}e[B_i|3, 2].
\end{aligned}$$

Notemos que la distribución final da mayor probabilidad a $\mathcal{B}e(2, 2)$, ya que si $Y = 1$, es más probable que θ provenga de la distribución $\mathcal{B}e(2, 2)$ en la mezcla inicial. Este es un ejemplo específico de una propiedad de las mezclas de procesos Dirichlet que se establece formalmente en el Corolario 2.4.

Ahora, establecemos dos corolarios del Teorema 2.4 que trata casos que ocurren frecuentemente en las aplicaciones.

Corolario 2.3 *Sean $\mathcal{P}$ un proceso Dirichlet en un espacio Borel estándar $(\Theta, \mathcal{A})$ con parámetro α, y θ una muestra aleatoria de una realización de $\mathcal{P}$. Sean $(\mathcal{Y}, \mathcal{C})$ un espacio de Borel estándar y F una probabilidad de transición en $\mathcal{C} \times \Theta$ al intervalo $[0, 1]$. Si la distribución condicional de Y dado θ es $F(\cdot; \theta)$, entonces la distribución condicional de $\mathcal{P}$ dado $Y=y$ es una mezcla de procesos Dirichlet en $(\Theta, \mathcal{A})$ con distribución de mezcla H en el espacio índice $(\Theta, \mathcal{A})$ y medida de transición $\alpha(x, \cdot) = \alpha(\cdot) + \delta_x(\cdot)$, donde la distribución de mezcla H en $(\Theta, \mathcal{A})$ es la distribución condicional de θ dado $Y = y$; esto es,*

$$\mathcal{P} = \mathcal{DP}(\alpha), \quad \theta \sim G,\ G \sim \mathcal{P}, \quad Y|\theta \sim F(\cdot; \theta) \ \Rightarrow\ \mathcal{P}|Y = \int \mathcal{DP}(\alpha + \delta_\theta)\, dH_y(\theta).$$

Es decir, si una muestra tomada de una realización de un simple proceso Dirichlet es distorsionada por un error aleatorio, entonces la distribución final del proceso es una mezcla de procesos Dirichlet. La demostración de este corolario es inmediata, tomando a $\mathcal{U}$ como un espacio con un sólo elemento en el Teorema 2.5.

COROLARIO 2.4 *Sea $\mathcal{P}$ una mezcla de procesos Dirichlet en un espacio de Borel estándar $(\Theta, \mathcal{A})$ con espacio índice de Borel $(\mathcal{U}, \mathcal{B})$, distribución H en $(\mathcal{U}, \mathcal{B})$ y medida de transición α en $\Theta \times \mathcal{A}$. Si θ es una muestra aleatoria de una realización de $\mathcal{P}$, entonces $\mathcal{P}$ dado θ es una mezcla de procesos Dirichlet en $(\Theta, \mathcal{A})$ con medida de transición $\alpha + \delta_\theta$ y distribución H_θ en $(\mathcal{U}, \mathcal{B})$, donde H_θ es la distribución condicional de u dado θ. Esto es, si $\mathcal{P} = \int_{\mathcal{U}} \mathcal{DP}(\alpha_u)\, dH(u)$, $\theta \sim G$ y $G \sim \mathcal{P}$, entonces $(\mathcal{P}|\theta) = \int_{\mathcal{U}} \mathcal{DP}(\alpha_u + \delta_\theta)\, dH_\theta(u)$.*

El punto esencial de este corolario es que la observación θ afecta a cada componente de la mezcla como esperaríamos que lo hiciera, añadiendo δ_θ a α_u. Además, cambia los pesos relativos de sus componentes a la distribución condicional de u dado θ. La demostración del corolario es inmediata del Teorema 2.5 degenerando la distribución de Y en θ.

Antoniak (1974) discute una generalización de este último resultado para el caso de una muestra aleatoria de una realización de una mezcla de procesos Dirichlet de tamaño n.

3. APLICACIONES DE LAS MEZCLAS BASADAS EN PROCESOS DIRICHLET

3.1 *Planteamiento*

Si tenemos una muestra aleatoria $\theta_1, \theta_2, \ldots, \theta_n$ cuya distribución G no conocemos, pero de la cual tenemos la idea de que, en su forma, es parecida a una función de distribución G_0, entonces podemos hacer inferencias sobre la población suponiendo que la muestra proviene de un proceso Dirichlet con parámetro $\alpha = \alpha G_0$, donde, como se verá más adelante, α es una constante que representa el grado de credibilidad de qué tanto la forma de G es parecida a la de G_0. El inconveniente de la solución anterior es que de acuerdo al Corolario 2.1, $\theta_1, \theta_2, \ldots, \theta_n$ provendrían de una medida de probabilidad discreta, es decir, se está asumiendo que la probabilidad de que haya datos repetidos es positiva.

Dada esta limitación en la búsqueda de un modelo Bayesiano no paramétrico que tenga como distribución inicial a una medida de probabilidad que asigne medida 1 al espacio de las distribuciones continuas, se han propuesto, al menos, dos posibles soluciones: la utilización de árboles de Pólya (Lavine 1992, 1994), y el uso de mezclas basadas en procesos Dirichlet. El inconveniente de la primera propuesta radica en la dificultad de su implementación (en particular, en el hecho de que la medida aleatoria resultante depende de la partición de Θ que se elija), mientras que, en contraste, la segunda es fácil de trabajar aprovechando la estructura jerárquica de las mezclas basadas en procesos Dirichlet. Esta segunda alternativa será la que revisaremos en este trabajo.

El modelo jerárquico típico es el siguiente: tenemos un conjunto de n observaciones $y_1, y_2, \ldots, y_n$ modeladas como provenientes, de manera independiente, de distribuciones $F_i(y_i|\theta_i, \sigma)$ conocidas para cada i, donde, θ_i es el conjunto de parámetros específicos para cada caso y σ el de parámetros comunes. Suponemos que la distribución de $\theta_1, \theta_2, \ldots, \theta_n$ es G y la de σ es $p(\sigma)$, donde, a su vez, G y $p(\sigma)$ pueden depender de hiperparámetros.

La manera natural de introducir el concepto de mezclas de procesos Dirichlet en este modelo es suponiendo incertidumbre en G y asumiendo que dicha distribución proviene de un proceso Dirichlet $\mathcal{P}$ con parámetro $\alpha G_0(\cdot|\gamma)$ donde $G_0(\cdot|\gamma)$ es una función de distribución continua y α, así como los hiperparámetros de G_0, γ, son variables aleatorias,

generalmente independientes, con distribuciones especificadas por $p(\alpha)$ y $p(\gamma)$ respectivamente, es decir,

$$\begin{aligned} &I. && y_i \sim F_i(y_i|\theta_i, \sigma) \text{ para } i = 1, \ldots, n. \\ &II. && \theta_1, \ldots, \theta_n \overset{i.i.d}{\sim} G. \\ &III. && G \sim \mathcal{DP}(\alpha G_0(\cdot|\gamma)). \\ &IV. && \alpha \sim p(\alpha),\ \gamma \sim p(\gamma) \text{ y } \sigma \sim p(\sigma). \end{aligned}$$

Lo anterior implica que estamos suponiendo que $\theta_1, \ldots, \theta_n$ es una muestra aleatoria de una realización de una mezcla de procesos Dirichlet con medida de transición $\alpha G_0(\cdot|\gamma)$.

Por el Corolario 2.1, sabemos, que a pesar de que G_0 es una distribución continua, G es discreta, es decir, la probabilidad de tener valores de θ repetidos es positiva.

Los niveles I a III (con α y γ fijas y donde σ sí puede ser una variable aleatoria) corresponden al modelo de Lo (1984), mientras que los niveles II a IV corresponden esencialmente al de Antoniak (1974). Escobar y West (1995) unen los dos modelos anteriores incluyendo todos los niveles (del I al IV), y es éste al que llamaremos una mezcla basada en procesos Dirichlet (cada y_i sigue una distribución de mezcla basada en procesos Dirichlet) y la función de distribución de y_i en el primer nivel (F_i) determinará si dicha variable es continua o discreta.

La intención de implementar una simulación por medio de un muestreo de Gibbs en los cálculos requeridos para estimar la distribución de observaciones futuras de y, hace necesario encontrar la distribución de una muestra aleatoria de tamaño 1 de una realización de $\mathcal{P}$ dada una de tamaño $n-1$ e $y_1, y_2, \ldots, y_n$. Por el momento, consideraremos los parámetros α, γ y σ conocidos y, por ello, los suprimiremos de la notación.

Dada una muestra aleatoria $\theta_1, \theta_2, \ldots, \theta_n$ de una realización de $\mathcal{P}$, definimos θ^{-i} como la muestra aleatoria después de extraer el valor θ_i, es decir, $\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n$. Por el Teorema 2.1, $\mathcal{P}|\theta^{-i} = \mathcal{DP}(\alpha + \sum_{j \neq i} \delta_{\theta_j})$, donde $\alpha = \alpha G_0(\cdot|\gamma)$. Esto implica, por la Proposición 2.4, que

$$\mathbb{P}[\theta_i \in A|\theta^{-i}] = \frac{\alpha}{\alpha + n - 1} G_0[A] + \frac{1}{\alpha + n - 1} \sum_{j \neq i} \delta_{\theta_j}(A), \tag{3.1}$$

donde por $G_0[A]$ entendemos la probabilidad que asigna la función de distribución G_0 al conjunto A: $\int_A g_0(x)\, dx$, denotando a la función de densidad correspondiente a G_0 como g_0.

La expresión (3.1) nos indica que la probabilidad de que θ_i sea distinta de θ_j, para $j \neq i$, es $\alpha/(\alpha + n - 1)$; además, que si α es pequeña entonces G tiende a concentrarse en átomos y, por el contrario, si α es grande, el modelo no paramétrico está "cercano a G_0", lo que confirma que α puede pensarse como el grado de credibilidad de qué tanto la forma de G es parecida a la de G_0. También implica que,

$$p(\theta_i|\theta^{-i}) = \frac{\alpha}{\alpha + n - 1} g_0(\theta_i) + \frac{1}{\alpha + n - 1} \sum_{j \neq i} \delta_{\theta_j}(\theta_i).$$

Usando la ecuación anterior, asumiendo la independencia de $y_i|\theta_i$ e $y_j|\theta_j$, y aplicando el Teorema de Bayes obtenemos:

$$\begin{aligned} p(\theta_i|\theta^{-i}, y_1, \ldots, y_n) &\propto p(y_1, y_2, \ldots, y_n|\theta_i, \theta^{-i}) p(\theta_i|\theta^{-i}) \\ &\propto \prod_{j=1}^{n} p(y_j|\theta_j) \Big\{ \alpha g_0(\theta_i) + \sum_{j \neq i} \delta_{\theta_j}(\theta_i) \Big\} \\ &\propto \alpha p(y_i|\theta_i) g_0(\theta_i) + \sum_{j \neq i} p(y_i|\theta_j) \delta_{\theta_j}(\theta_i). \end{aligned}$$

Observando que,

$$p(y_i|\theta_i) g_0(\theta_i) = \frac{p(y_i|\theta_i) g_0(\theta_i)}{p(y_i)} p(y_i) = p(\theta_i|y_i) \int p(y_i|\theta) g_0(\theta)\, d\theta,$$

obtenemos,

$$p(\theta_i|\theta^{-i}, y_1, y_2, \ldots, y_n) \propto q_{i,0} g_{i,0}(\theta_i) + \sum_{j \neq i} q_{i,j} \delta_{\theta_j}(\theta_i), \tag{3.2}$$

donde $q_{i,0} = \alpha \int p(y_i|\theta) g_0(\theta)\, d\theta$, es decir, el producto de α y la distribución marginal de $p(y_i)$, $q_{i,j} = p(y_i|\theta_j)$ y $g_{i,0}(\theta_i) \propto p(y_i|\theta_i) g_0(\theta_i)$, la distribución de θ_i dado y_i.

Con la finalidad de hacer una simulación más eficiente, definimos la configuración de una realización de $\mathcal{P}$, $\theta_1, \theta_2, \ldots, \theta_n$, como el número de valores distintos en la muestra, que llamaremos k, dichos valores, $\theta^* = \{\theta_1^*, \theta_2^*, \ldots, \theta_k^*\}$ y el vector de funciones indicadoras $\varphi = (\varphi_1, \varphi_2, \ldots, \varphi_n)$ donde $\varphi_i = j$ si $\theta_i = \theta_j^*$. Notemos que conocer los valores de la una realización de $\mathcal{P}$ equivale a conocer los valores de su configuración.

De esta manera, dicha configuración (West 1990, MacEachern 1994) determina una forma de clasificar los datos $Y = \{y_1, \ldots, y_n\}$ en k grupos diferentes con $n_j = \{i|\varphi_i = j\}$ observaciones que comparten el parámetro θ_j^* en cada grupo j, $j = 1, \ldots, k$. Denotaremos con I_j al conjunto de índices de las observaciones del grupo j, $I_j = \{i|\varphi_i = j\}$ e

$Y_j = \{y_i|\varphi_i = j\}$ al correspondiente conjunto de observaciones. En el muestreo de Gibbs se simularán, en un primer paso, una configuración dada la anterior y, en un segundo, $\theta_1^*, \theta_2^*, \ldots, \theta_k^*$ dadas φ y k.

Por otro lado, como las $\theta_j^{*\prime}s$ son una muestra aleatoria de la distribución inicial G_0, el análisis posterior lleva a una colección de k análisis independientes; específicamente, las $\theta_j^{*\prime}s$ son, entre ellas, condicionalmente independientes con densidades finales:

$$\begin{aligned} p(\theta_j^*|Y, \varphi, k) &\propto p(Y|\theta_j^*, \varphi, k)p(\theta_j^*|\varphi, k) \\ &\propto \Big\{ \prod_{i \in I_j} f_i(y_i|\theta_j^*) \Big\} g_0(\theta_j^*). \end{aligned} \tag{3.3}$$

Ahora bien, si denotamos por k^{-i}, n_j^{-i} e I_j^{-i} para $j = 1, \ldots, k^{-i}$, $\theta^{*-i} = (\theta_1^{*-i}, \ldots, \theta_{k^{-i}}^{*-i})$ a la configuración correspondiente a la muestra aleatoria θ^{-i}, al sustituir en (3.1) obtenemos,

$$\mathbb{P}[\theta_i \in A|\theta^{-i}] = \frac{\alpha}{\alpha + n - 1} G_0[A] + \frac{1}{\alpha + n - 1} \sum_{j=1}^{k^{-i}} n_j^{-i} \delta_{\theta_j^*}(A), \tag{3.4}$$

Esto muestra que θ_i es distinto de los otros parámetros y generado de G_0 con probabilidad $\alpha/(\alpha + n - 1)$, de otra manera es elegida de los valores ya observados $\theta_j^{*-i\prime}s$ de acuerdo a una distribución multinomial con parámetros proporcionales a la cantidad de elementos de cada grupo I_j^{-i}, n_j^{-i}.

La extensión de (3.4) de n a $n+1$ es relevante al momento de estimar la distribución final de G dado $\theta_1, \ldots, \theta_n$, y así predecir el nuevo valor de θ para el caso $i = n + 1$. Incidentalmente, dicha distribución es la esperanza de G dados los valores de $\theta_1, \ldots, \theta_n, \varphi, k$, es decir,

$$\mathbb{P}(\theta_{n+1} \in A|\theta_1, \ldots, \theta_n, \varphi, k) = E(G|\theta_1, \ldots, \theta_n, \varphi, k) = \frac{\alpha}{\alpha + n} G_0[A] + \frac{1}{\alpha + n} \sum_{j=1}^{k} n_j \delta_{\theta_j}(A). \tag{3.5}$$

Como resultado de la ecuación anterior, la distribución predictiva de la observación futura y_{n+1} es

$$y_{n+1}|\theta^*, \varphi, k \sim \frac{\alpha}{\alpha + n} F_{n+1}(\cdot|\theta_{n+1}) + \frac{1}{\alpha + n} \sum_{j=1}^{k} n_j F_{n+1}(\cdot|\theta_j^*), \tag{3.6}$$

donde θ_{n+1} es una nueva muestra de G_0 y F_{n+1} es la función de distribución correspondiente a y_{n+1}. Esta distribución predictiva podrá utilizarse como estimador de la densidad

desconocida de y (los datos). Un ejemplo de dicha aplicación se encuentra en la Sección 4.5.

3.2 *Muestreo de Gibbs para mezclas basadas en procesos Dirichlet*

MacEachern (1994), propuso el siguiente algoritmo de muestreo de Gibbs para simular muestras aleatorias $\theta_1, \ldots, \theta_n$ de una realización de $\mathcal{P}|Y$ trabajando en términos de los parámetros equivalentes (configuraciones) k, θ^* y φ (nótese que se está suponiendo que σ y α son conocidos; la simulación de dichos parámetros puede agregarse fácilmente al esquema del muestreo de Gibbs, como se verá más adelante, y por ello, dichas variables las mantenemos fuera de la notación).

Reescribiendo (3.2),

$$p(\theta_i|Y, \theta^{-i}, \varphi^{-i}, k^{-i}) = q_{i,0} g_{i,0}(\theta_i) + \sum_{j=1}^{k^{-i}} q_{i,j} \delta_{\theta_j^{*-i}}(\theta_i), \tag{3.7}$$

donde, ahora, los pesos $q_{i,j}$ están dados por,

$$q_{i,j} = \begin{cases} c\alpha h_i(y_i), & \text{si } j = 0, \\ c n_j^{-i} f_i(y_i|\theta_j^*), & \text{si } j \geqslant 1, \end{cases}$$

donde f_j es la función de densidad correspondiente a F_j, $g_{i,0}$ es la función de densidad final de θ, cuya correspondiente función de distribución llamaremos $G_{i,0}$, con densidad obtenida de la actualización de g_0 vía la función de densidad $f_i(y_i|\theta_i)$, es decir,

$$g_{i,0}(\theta_i) \propto f_i(y_i|\theta_i) g_0(\theta_i),$$

cuya constante de normalización, $h_i(y_i)$ es la densidad marginal de y_i,

$$h_i(y_i) = \int f_i(y_i|\theta) g_0(\theta)\, d\theta,$$

y c es una constante de normalización.

La ecuación (3.7) implica inmediatamente distribuciones finales para la configuración de variables indicadoras,

$$\mathbb{P}(\varphi_i = j|Y, \theta^{-i}, \varphi^{-i}, k^{-i}) = q_{i,j}. \tag{3.8}$$

Ahora podemos simular muestras de valores de k, φ y $\theta^* = \{\theta_1^*, \ldots, \theta_k^*\}$ iterando de la siguiente manera:

(a) Dados valores iniciales de k, θ^* y φ, generamos una nueva configuración simulando sucesivamente valores de los indicadores a partir de la distribución final (3.8), remplazando $\varphi_1, \varphi_2, \ldots$; para cualquier índice i tal que $\varphi_i = 0$ obtenemos una nueva muestra de $G_{i,0}$ en (3.7).

Si bien iteraciones sucesivas del paso (a) convergen, eventualmente, a valores aproximados de la distribución final $p(k, \theta^*, \varphi|Y)$ produciendo una cadena de Markov ergódica, dicha convergencia es muy lenta, y, por ende, el muestreo es ineficiente. El problema es que pueden haber varios grupos de observaciones que, con alta probabilidad, estén asociadas al mismo valor de θ. El algoritmo anterior no cambia el valor de θ para más de una observación simultáneamente, por lo que el cambio de los valores de θ en tales grupos ocurre pocas veces, ya que para dicho cambio la cadena requiere pasar por un estado intermedio de baja probabilidad en el cual las observaciones del grupo no estén asociadas al mismo valor de θ. Para evitar la situación anterior, es necesario que, una vez obtenidos los valores de k, θ^*, y φ en el paso (a), se remuestreen los valores de θ^* dados k y φ, es decir, se debe agregar al algoritmo (después de cada paso (a)),

(b) Dados k y φ, obtenemos un nuevo conjunto de parámetros θ^* generando nuevos valores de la distribución final en (3.3) proporcional a $\left\{ \prod_{i \in I_j} f_i(y_i|\theta_j^*) \right\} g_0(\theta_j^*)$.

Los temas relativos a la convergencia de la sucesión de valores de k, θ^* y $\varphi|Y$ se discuten en MacEachern y Müller (1998). Las inferencias pueden basarse en histogramas de valores muestreados para parámetros individuales. Por ejemplo, el promedio de (3.6), con respecto a valores simulados de $\theta_1^*, \ldots, \theta_k^*, \theta_{n+1}$, provee un estimador de la función de densidad predictiva $p(y_{n+1}|Y)$. Tal ejemplo está basado en aproximaciones de valores $\theta_1^*, \ldots, \theta_k^*$ generados a partir de una iteración de Gibbs y valores θ_{n+1} generados directamente de G_0.

Todo lo mencionado anteriormente en esta capítulo está condicionado al parámetro común σ. En algunos modelos dicho parámetro no será considerado y no habrá más que hacer. Si, por el contrario, σ es incluido en el modelo, entonces deberemos extender el muestreo de Gibbs con la finalidad de incluir los pasos necesarios para generar valores de

σ de la distribución condicional final apropiada de la siguiente forma: dada $p(\sigma)$, tenemos

$$\begin{aligned} p(\sigma|Y,\theta^*,\varphi,k) &\propto p(Y|\theta^*,\varphi,k,\sigma)p(\sigma|\theta^*,\varphi,k) \\ &\propto p(\sigma)\prod_{i=1}^{n} f_i(y_i|\theta_i,\sigma) \\ &\propto p(\sigma)\prod_{j=1}^{k}\prod_{i\in I_j} f_i(y_i|\theta_j^*,\sigma) \end{aligned} \tag{3.9}$$

El esquema de muestreo dado anteriormante, se extiende, si es necesario, agregando el paso (b*) que consiste en:

(b*) Incluir valores simulados de σ, condicionando en los valores del resto de los parámetros recientemente muestreados, θ^*, φ y k a partir de la expresión (3.9), en cada paso.

Retomando, el hiperparámetro de G_0, que hemos denominado γ, por lo general será incierto pero con una distribución inicial especificada.

Las distribuciones (3.3) y (3.8) de los pasos (a) y (b) son condicionados a $\pi = \{\alpha, \gamma\}$. El muestreo de Gibbs se extiende directamente para incorporar π simplemente agregando un paso más al muestreo de la distribución final apropiada $p(\pi|Y,\theta,\sigma) = p(\pi|Y,\theta^*,\varphi,k,\sigma)$. Esto puede desarrollarse como lo hicieron West (1992) y Escobar y West (1995) de la siguiente manera:

Primero asumimos que α y γ son independientes con densidades específicas

$$p(\alpha,\gamma) = p(\alpha)p(\gamma).$$

Entonces, dada la estructura del modelo, α y γ siguen siendo condicionalmente independientes dados los parámetros Y, θ^*, φ, k, y σ; por lo que α y γ pueden ser consideradas por separado.

En efecto, por la estructura del proceso Dirichlet (hecho resaltado en (2.17), donde se ve que la distribución de k sólo depende de $\alpha = \alpha(\Theta)$), únicamente k es relevante en la inferencia de α, por lo que

$$p(\alpha|Y,\theta^*,\varphi,k,\sigma) = p(\alpha|k). \tag{3.10}$$

Las aplicaciones de hoy en día, asumen una mezcla de distribuciones gamma para la inicial $p(\alpha)$, que lleva nuevamente a una mezcla de gammas a $p(\alpha|k)$ (West, 1992), o bien se toma a $p(\alpha)$ simplemente como una distribución gamma que implica, como veremos en la siguiente sección, una simulación fácil de $p(\alpha|k)$.

Para γ, notando que dicho parámetro entra al modelo sólo a través de G_0, obtenemos, a partir de (3.3),

$$p(\gamma|Y,\theta^*,\varphi,k,\sigma) = p(\gamma|\theta^*,k) \propto p(\gamma)\prod_{j=1}^{k} g_0(\theta_j^*|\gamma). \tag{3.11}$$

Distribuciones iniciales de G_0 que permitan un muestreo directo de la distribución posterior (3.11) deben de tomarse en cuenta al momento de las aplicaciones. Asumiendo que este es el caso, el conocimiento de Y,θ^*,φ,k y σ se reduce a la información de θ^* y k, lo que lleva a la simulación de valores de $\pi = \{\alpha,\gamma\}$:

(c) Dada k, obtengamos una muestra de $p(\alpha|k)$ (West 1992, y Escobar y West 1995), dado k y φ conseguimos un nuevo valor de γ de (3.10).

Los valores resultantes para α y γ son usados condicionalmente en (a) y (b) para la iteración siguiente.

3.3 *Cálculo y simulación de* $p(\alpha|k)$

A continuación, veremos que la elección de una familia específica pero flexible de distribuciones iniciales para α nos lleva a una simulación sencilla de $p(\alpha|k)$.

La expresión (2.17) puede ser rescrita como

$$p(k|\alpha,n) = c_n(k)n!\alpha^k\frac{\Gamma(\alpha)}{\Gamma(\alpha+n)}, \quad (k=1,2,\ldots,n), \tag{3.12}$$

donde $c_n(k) = \mathbb{P}(k|\alpha=1,n)$, que no involucra a α y que, si se requiere, puede ser calculado utilizando los números de Stirling de primera clase.

Supongamos una distribución inicial continua para $p(\alpha)$ (que a su vez puede depender del tamaño de muestra n pero que, por claridad, suprimiremos de la notación), lo que implica que $p(k) = \int p(k|\alpha)p(\alpha)\,d\alpha$. Supongamos también que ya hemos generado al parámetro k y al resto de los parámetros del modelo. Por la expresión (3.10),

$$p(\alpha|k,Y,\theta) \propto p(\alpha|k) \propto p(\alpha)p(k|\alpha), \tag{3.13}$$

con función de verosimilitud dada por (3.12); entonces el muestreo de Gibbs puede ser extedido para considerar a α; para α dada generamos al resto de los parámetros y en cada iteración, generamos α a partir del valor de k. Una forma sencilla que desarrollaremos a

continuación para muestrear de (3.13) es posible cuando la distribución inicial $p(\alpha)$ es una distribución gamma con parámetro de forma a y de escala b, es decir $\alpha \sim \mathcal{G}(\alpha|a,b)$ que tiene como función de densidad $p(\alpha) = (1/\Gamma(a))b^a \exp(-\alpha b)\alpha^{a-1}1_{(0,\infty)}(\alpha)$.

La función gamma en (3.12) puede escribirse como,

$$\frac{\Gamma(\alpha)}{\Gamma(\alpha+n)} = \frac{(\alpha+n)\mathrm{B}(\alpha+1,n)}{\alpha\Gamma(n)},$$

donde, $\mathrm{B}(u,v) = \int_0^1 x^{u-1}(1-x)^{v-1}\,dx = \Gamma(u)\Gamma(v)/\Gamma(u+v)$ para $u,v>0$, es la función beta usual. Entonces (3.13), para $k = 1,\ldots,n$, puede escribirse como,

$$\begin{aligned} p(\alpha|k) &\propto p(\alpha)\alpha^{k-1}(\alpha+n)\mathrm{B}(\alpha+1,n) \\ &\propto p(\alpha)\alpha^{k-1}(\alpha+n)\int_0^1 x^{\alpha}(1-x)^{n-1}\,dx. \end{aligned}$$

La expresión anterior implica que $p(\alpha|k)$ es la marginal de la distribución conjunta de α y una cantidad continua x $(0<x<1)$ tal que,

$$p(\alpha|x,k) \propto p(\alpha,x|k) \propto p(\alpha)\alpha^{k-1}(\alpha+n)x^{\alpha}(1-x)^{n-1}, \qquad (0<\alpha, 0<x<1),$$

entonces, la distribución condicional final $p(\alpha|x,k)$ está dada por,

$$\begin{aligned} p(\alpha|x,k) &\propto p(\alpha,x|k) \\ &\propto p(\alpha)\alpha^{k-1}(a+n)x^{\alpha} \\ &\propto \{e^{-\alpha b}\alpha^{a-1}\}\alpha^{k-1}(a+n)x^{\alpha} \\ &\propto \{\alpha^{a+k-2}e^{-\alpha b+\alpha\log x}\}(\alpha+n) \\ &\propto \alpha^{a+k-1}e^{-\alpha(b-\log x)} + n\alpha^{a+k-2}e^{-\alpha(b-\log x)} \\ &\propto \pi_x\mathcal{G}(\alpha|a+k,b-\log x) + (1-\pi_x)\mathcal{G}(\alpha|a+k-1,b-\log x), \end{aligned} \qquad (3.14)$$

donde,

$$\frac{\pi_x}{1-\pi_x} = \frac{\Gamma(a+k)/(b-\log x)^{a+k}}{\Gamma(a+k-1)/(n(b-\log x)^{a+k-1})} = \frac{a+k-1}{n(b-\log x)},$$

(notemos que esta distribución está definida para cualquier distribución inicial gamma y x en el intervalo unitario).

Ahora,

$$p(x|\alpha,k) \propto p(\alpha,x|k) \propto x^{\alpha}(1-x)^{n-1}, \qquad (3.15)$$

esto es, $x|\alpha,k$ se distribuye beta con parámetros $\alpha+1,n$, es decir, $x|\alpha,k \sim \mathcal{B}e(x|\alpha+1,n)$.

Entonces α puede ser generado en cada paso de la simulación -en cada iteración del muestreo de Gibbs- una vez simulado el valor de k, primero (i) se genera x de una simple distribución beta (3.15), condicionado en el valor más reciente de α, y posteriormente (ii) se simula el nuevo valor de α de la mezcla de gammas en (3.14) basada en el mismo valor de k y el valor de x generado en (i).

Para completar la simulación, tendremos una serie de valores k, α, x y el resto de los parámetros. Supongamos que el tamaño de muestra de Monte Carlo es N, y denotemos los valores generados k_s, x_s, etc., para $s = 1, \ldots, N$. Únicamente los valores k_s, x_s se necesitan para estimar la distribución final de α vía el promedio de Monte Carlo de densidades condicionales finales, es decir,

$$p(\alpha|Y) = \frac{1}{N}\sum_{s=1}^{N} p(\alpha|x_s, k_s). \tag{3.16}$$

3.4 *Ejemplo de una aplicación de mezclas basadas en procesos Dirichlet utilizando distribuciones iniciales conjugadas*

Gelfand y Smith (1990) estudiaron la curva de incremento de peso de dos grupos (uno de control y otro después de un tratamiento) de individuos (ratas) utilizando una regresión lineal con errores normales. Las distribuciones iniciales de los coeficientes de regresión se supusieron, como usualmente se hace, normales y se consideró que el incremento de peso de cada individuo era lineal durante el periodo del experimento. El planteamiento del análisis es el siguiente:

Denotemos con $y_{i,j}$ al peso de la rata i en la semana j ($n = 30$, $i = 1, \ldots, n$, $j = 1, \ldots, 5$), entonces suponemos que,

$$(y_i|\mu_i, \sigma) \sim N(y_i|X\mu_i, \sigma^2 I), \qquad \text{para } i = 1, \ldots, n \tag{3.17}$$

donde $y_i = (y_{i,1}, \ldots, y_{i,5})^t$, I es la matriz identidad de 5×5, σ^2 es el parámetro común a todos los individuos y X es una matriz de 5×2 cuya j-ésima columna es $(1, x_j)$ y x_j es la edad en días de las ratas al momento de la j-ésima medición, es decir,

$$X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 8 & 15 & 22 & 29 \end{pmatrix}.$$

Gelfand y Smith (1990) utilizaron distribuciones iniciales conjugadas para los parámetros en (3.17) proponiendo el siguiente modelo:

$$\mu_i|m,B \sim N(\mu_i|m,B), m \sim N(m|a,A), B \sim W(B|2c,(cC)^{-1}) \text{ y } \sigma^2 \sim \mathcal{GI}(\sigma^2|u/2,uU/2),$$

donde $a,c,u,U \in \mathbb{R}$, $c,u,U \geqslant 0$, A y C dos matrices de varianzas de 2×2, $\mathcal{GI}(\cdot|a,b)$ denota la distribución gamma inversa y $W(\Sigma|s,S)$ a la Wishart de dimensión p con esperanza sS,

$$W(\Sigma|s,S) \propto |S|^{\frac{-s}{2}}\Sigma^{\frac{s-p-1}{2}}\exp(-traza(S^{-1}\Sigma)/2). \tag{3.18}$$

Para la simulación a través del muestreo de Gibbs se calcularon las siguientes distribuciones de cada parámetro condicionado al resto de ellos y se hacen explícitas sólo las cantidades relevantes (por ejemplo, como $\sigma^2|Y,\mu,m,B$ no depende ni de m ni de B sólo escribimos $\sigma^2|Y,\mu$):

$$\begin{aligned}
(\sigma^2|Y,\mu) &\sim \mathcal{GI}\left(\sigma^2\left|\frac{u+5n}{2},\frac{1}{2}\left[uU+\sum_{i=1}^{n}(y_i-X\mu_i)^t(y_i-X\mu_i)\right]^{-1}\right.\right)\\
(\mu_i|Y,m,B,\sigma) &\sim N(\mu_i|D(B^{-1}m+\sigma^{-2}X^t y_i),D)\\
(m|\mu,B) &\sim N\left(m\left|V\left(A^{-1}a+B^{-1}\sum_{i=1}^{n}\mu_i\right),V\right.\right)\\
(B^{-1}|m,\mu) &\sim W\left(B^{-1}\left|c+n,\left[cC+\sum_{i=1}^{n}(\mu_i-m)(\mu_i-m)^t\right]^{-1}\right.\right),
\end{aligned} \tag{3.19}$$

donde $D=(B^{-1}+\sigma^{-2}X^tX)^{-1}$ y $V=(A^{-1}+nB^{-1})^{-1}$.

Se usaron distribuciones iniciales difusas para σ^2, m y B dando a los hiperparámetros los siguientes valores:

$$u=0, \qquad A=0, \qquad c=2 \text{ y} \qquad C=\begin{pmatrix}100 & 0\\ 0 & 0.1\end{pmatrix}. \tag{3.20}$$

West et al. (1994) plantearon el problema anterior desde un punto de vista no paramétrico introduciendo incertidumbre en la distribución de μ_i, que llamaremos θ_i para seguir con la notación de la sección anterior, y dándole a estas variables una distribución G proveniente de un proceso Dirichlet con parámetro $\alpha N(\cdot|m,B)$, es decir, suponiendo que si bien, la distribución θ_i para $i=1,\ldots,n$ no es normal si está "cercana a ella" (se

tomaron los datos del grupo de control del artículo de Gelfand y Smith, 1990). El modelo que plantean es el siguiente:

$$
\begin{aligned}
&I. && y_i|\theta_i, \sigma^2 \sim N(y_i|X\theta_i, \sigma^2 I), \text{ para } i = 1, \ldots, n, \\
&II. && \theta_1, \ldots, \theta_n \overset{i.i.d}{\sim} G, \\
&III. && G|m, B \sim \mathcal{DP}(\alpha N(\cdot|m, B)), \\
&IV. && m \sim N(m|a, A), \quad B \sim W(B|2c, (cC)^{-1}), \\
& && \sigma^2 \sim \mathcal{GI}(\sigma^2|u/2, uU/2), \text{ y } \alpha \sim \mathcal{G}(\alpha|1/2, 1/2),
\end{aligned}
\tag{3.21}
$$

y u, A, c y C tomaron los mismos valores que en el modelo anterior (los dados en (3.20)).

Comparando la notación de Gelfand y Smith (1990) y las de las secciones anteriores de este capítulo, $y_i, \theta_i, \sigma^2, (m, B), \alpha$ y $N(\theta_i|m, B)$ juegan el papel de $y_i, \theta_i, \sigma^2, \gamma, \alpha$ y $G_0(\theta_i|\gamma)$ respectivamente. Asimismo, para implementar el muestreo de Gibbs en el nuevo modelo se necesitan las variables θ^*, φ, k y los conjuntos de índices I_j con cardinalidad n_j para $j = 1, \ldots, k$, y también, $\theta^{-i*}, \varphi^{-i}, k^{-i}$ y los conjuntos de índices $, I_j^{-i}$ con cardinalidad n_j^{-i} para $j = 1, \ldots, k^{-i}$ que son las configuraciones correspondientes a $\theta = (\theta_1, \ldots, \theta_n)$ y $\theta^{-i} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n)$ definidas en la Sección 3.1 respectivamente.

Una vez dados los valores iniciales, para $k, \theta^* = (\theta_1^*, \ldots, \theta_k^*), \varphi = (\varphi_1, \ldots, \varphi_k), m, B, \sigma, \alpha$ los pasos (a), (b) y (c) de la sección anterior, para este problema en particular, se traducen en:

(a') Simular $\theta_i|Y, \theta^{-i}$ para $i = 1, \ldots, n$ de la distribución:

$$
\theta_i|Y, \theta^{-i} \sim q_{i,0} G_{i,0}(\theta_i) + \sum_{j=1}^{k^{-i}} q_{i,j} \delta_{\theta_j^{*-i}}(\theta_i),
$$

donde, $G_{i,0}(\cdot) = N(\cdot|D[\sigma^{-2} X y_i + B^{-1} m], D)$, $D = (\sigma^{-2} X^t X + B^{-1})^{-1}$,

$$
\begin{aligned}
q_{i,0} &= c\alpha N(y_i|Xm, XBX^t + \sigma^2 I), \\
q_{i,j} &= cN(y_i|X\theta_j^*, \sigma^2 I), \qquad \text{para } j = 1, \ldots, k^{-i},
\end{aligned}
$$

I es la matriz identidad de 2×2 y c una constante de normalización (*i.e.*, tal que $\sum_{j=0}^{k^{-i}} q_{i,j} = 1$). Actualizando el nuevo valor de θ_i antes de simular θ_{i+1}, hasta, finalmente obtener nuevos valores para $\theta_1, \ldots, \theta_n$, o lo que es lo mismo, y más útil, una nueva configuración θ^*, φ, k.

(b') Una vez obtenidos los nuevos valores de φ y k, simulamos una nueva muestra de $\theta^* = (\theta_1^*, \ldots, \theta_k^*)$ de acuerdo a la distribución dada en (3.3), que, para este caso es,

$$(\theta_j^*|Y, \varphi, k) \sim N\left(\theta_j^* \middle| D_j^* \left[B^{-1}m + \sigma^{-2}X^t \sum_{i \in I_j} y_i\right], D_j^*\right),$$

donde $D_j^* = (B^{-1} + \sigma^{-2}n_j X^t X)^{-1}$.

En este caso sí se consideró un parámetro común a todas las observaciones, por lo que el paso (b*) se debe realizar, y se traduce en,

(b*') Simular σ de la distribución (3.9), que coincide con la ya dada en (3.19).

(c') Simular $\gamma = (m, B)$ de (3.11), que para este modelo es lo mismo que producir muestras de m y b de

$$(m|\theta^*, \varphi, k, B) \sim N\left(m \middle| V^*\left(A^{-1}a + B^{-1}\sum_{j=1}^{k}\theta_j^*\right), V^*\right)$$

$$(B^{-1}|\theta^*, \varphi, k, m) \sim W\left(B^{-1} \middle| c + k, \left[cC + \sum_{j=1}^{k}(\theta_j^* - m)(\theta_j^* - m)^t\right]^{-1}\right),$$

donde $V^* = (A^{-1} + kB^{-1})^{-1}$. Finalmente se simula α de (3.15), es decir, obtenemos una muestra x de,

$$(x|\alpha, k) \sim Be(x|\alpha + 1, n),$$

y luego α de

$$(\alpha|x, k) \sim \pi_x \mathcal{G}(\alpha|1/2 + k, 1/2 - \log x) + (1 - \pi_x)\mathcal{G}(\alpha|k - 1/2, 1/2 - \log x),$$

donde,

$$\pi_x = \frac{(k - 1/2)/(n(1/2 - \log x))}{1 + (k - 1/2)/(n(1/2 - \log x))} = \frac{k - 1/2}{n(1/2 - \log x) + k - 1/2}$$

En las Figuras 3.1 y 3.2 se muestran las distribuciones finales (histogramas suavizados) de μ_{n+1} en el modelo paramétrico de Gelfand y Smith (1990) y su equivalente θ_{n+1}, obtenido de muestras de la distribución dada en (3.5), en el no paramétrico de West et al. (1994). Aquí μ_{n+1} y θ_{n+1} denotan las medias de una observación futura correspondiente a una nueva rata. Mientras que el segundo modelo indica una bimodalidad de θ y por

ende, que la población de ratas es heterogénea, el primero, debido a la resticción inicial sobre la distribución de μ, señala que la distribución de μ es unimodal y no detecta la heterogeneidad de la población.

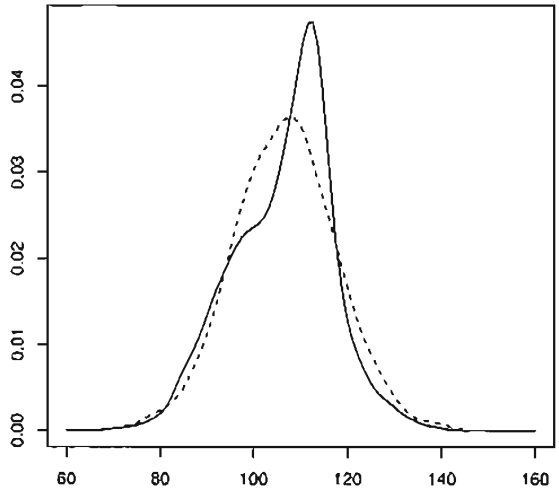

Figura 3.1 *Distribuciones finales de $\mu_{n+1,1}$ (linea punteada) y de $\theta_{n+1,1}$ (linea continua)*

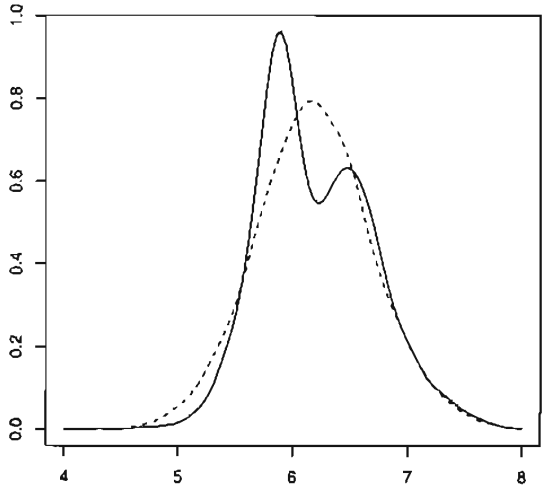

Figura 3.2 *Distribuciones finales de $\mu_{n+1,2}$ (linea punteada) y de $\theta_{n+1,2}$ (linea continua)*

3.5 *Ejemplo de una aplicación de mezclas basadas en Procesos Dirichlet en estimaciones de densidades multivariadas*

Supongamos que tenemos un conjunto $y_1, y_2, \ldots, y_n$ de observaciones cuya distribución no conocemos y que queremos estimar. Una manera de hacerlo es suponiendo el siguiente modelo jerárquico, en el que parametrizamos, para facilitar la notación, a la distribución normal en términos de su precisión en vez de su varianza:

Las observaciones $y_1, y_2, \ldots, y_n$ de dimensión p tienen cada una una distribución normal con media y matriz de precisión, en principio diferentes, es decir,

$$y_i \sim N(y_i|\mu_i, \lambda_i), \qquad \text{para } i = 1, \ldots, n.$$

Suponemos que los parámetros (μ_i, λ_i) son una muestra de una realización de una mezcla de procesos Dirichlet con las siguientes especificaciones:

$$(\mu_1, \lambda_1), \ldots, (\mu_n, \lambda_n) \overset{iid}{\sim} G \text{ y } G|\alpha, \gamma \sim \mathcal{DP}(\alpha G_0(\cdot|\gamma)),$$

donde, $G_0((\mu_i, \lambda_i)|m, T) \sim N(\mu_i|m, b\lambda_i)\mathcal{W}(\lambda_i|s, s^{-1}T)$, $m \sim N(m|a, A)$, $T^{-1} \sim \mathcal{W}(T^{-1}|s, s^{-1}I_p)$ y $\alpha \sim \mathcal{G}(\alpha|a_\alpha, b_\alpha)$ con b, s, a_α y b_α números reales positivos, a un vector en $\mathbb{R}^p$ y A una matriz simétrica y definida positiva de $p \times p$ e I_p la matriz identidad de $p \times p$. Es decir, de acuerdo con la notación dada al principio del capítulo, $N(\cdot|\mu_i, \lambda_i)$, (μ_i, λ_i), α, $N(\cdot|m, \cdot)$ $\mathcal{W}(\cdot\,\cdot|s, s^{-1}T)$ y (m, T) juegan, en este ejemplo, el papel de $F_i(\cdot|\theta_i), \theta_i, \alpha, G_0$ y γ respectivamente.

De acuerdo con el modelo propuesto, denotando con Y al conjunto $\{y_1, y_2, \ldots, y_n\}$, y con μ^* y λ^* a los conjuntos de k observaciones diferentes de $\{\mu_1, \ldots \mu_n\}$ y $\{\lambda_1, \ldots, \lambda_n\}$, $\{\mu_1^*, \ldots \mu_k^*\}$ y $\{\lambda_1^*, \ldots, \lambda_k^*\}$ respectivamente (es decir, $\{\mu^*, \lambda^*\}$ es lo que en la Sección 3.2 habíamos denominado por θ^*), y donde φ, el superíndice $-i$ e I_i se definen como al principio de este capítulo la distribución de (μ_i, λ_i) dada en la expresión (3.7) queda,

$$(\mu_i, \lambda_i)|Y, \mu^{-i}, \lambda^{-i}, \varphi^{-i}, k^{-i} \sim q_{i,0}G_{i,0}(\mu_i, \lambda_i) + \sum_{i=1}^{k^{-i}} q_{i,j}\delta_{(\mu_j^*, \lambda_j^*)}(\mu_i, \lambda_i), \tag{3.22}$$

donde,

$$q_{i,j} = \begin{cases} c\alpha h_i(y_i), & \text{si } j = 0, \\ cn_j^{-i} f_i(y_i|\mu_j^*, \lambda_j^*), & \text{si } j \geqslant 1, \end{cases}$$

$f_i(y_i|\mu_j^*, \lambda_j^*)$ es la función de densidad de una normal con media μ_j^* y matriz de precisión λ_j^* evaluada en y_i y al realizar las cuentas correspondientes, gracias a que estamos usando distribuciones conjugadas, obtenemos,

$$\begin{aligned} h_i(y_i) &= \int N(y_i|\mu_i, \lambda_i) N(\mu_i|m, b\lambda_i) W(\lambda_i|s, s^{-1}T)\, d(\mu_i, \lambda_i) \\ &= St_p\left(y_i \middle| m, \frac{(s-1)b}{s(b+1)}T, s-1\right), \end{aligned}$$

donde St_k denota la función de densidad Student k-variada, dada por,

$$\begin{aligned} &St_k(X|\mathrm{M}, \Lambda, \eta) = c[1 + \eta^{-1}(X - \mathrm{M})^t \Lambda (X - \mathrm{M})]^{-(\eta+k)/2}, \\ &c = \frac{\Gamma((\eta+k)/2)}{\Gamma(\eta/2)(\eta k)^{k/2}} |\Lambda|^{1/2}, \end{aligned}$$

$X, \mathrm{M} \in \mathbb{R}^k, \eta \in \mathbb{R}$ y Λ es una matriz simétrica y definida positiva de $k \times k$ y $G_{i,0}$ es la función de distribución con densidad,

$$\begin{aligned} g_{i,0}(\mu_i, \lambda_i) &\propto f_i(y_i|\mu_i, \lambda_i) g_0(\mu_i, \lambda_i) \\ &\propto N(y_i|\mu_i, \lambda_i) N(\mu_i|m, b\lambda_i) \mathcal{W}(\lambda_i|s, s^{-1}T) \\ &\propto N\left(\mu_i \middle| \frac{y_i + bm}{b+1}, (b+1)\lambda_i\right) \\ &\quad \times \mathcal{W}\left(\lambda_i \middle| s+1, \left[\frac{b}{b+1}(y_i - m)(y_i - m)^t + sT^{-1}\right]^{-1}\right). \end{aligned} \tag{3.23}$$

Por lo tanto, el paso (a) de la Sección 3.2 en este ejemplo indica que, dados los valores iniciales de $k, (\mu^*, \lambda^*)$ y φ, generamos una nueva configuración simulando de (3.8); para cualquier valor tal que $\varphi = 0$ obtenemos una nueva muestra de $G_{i,0}$ a partir de (3.22).

Procediendo de acuerdo al paso (b) de la Sección 3.2, dados los valores de k y φ se

genera un nuevo conjunto de valores (μ^*, λ^*) a partir de,

$$
\begin{aligned}
& p(\mu_j^*, \lambda_j^* | Y, m, T) \\
& \quad \propto \left\{ \prod_{i \in I_j} f_i(y_i | \mu_j^*, \lambda_j^*) \right\} g_0(\mu_j^*, \lambda_j^*) \\
& \quad \propto \left\{ \prod_{i \in I_j} N(y_i | \mu_j^*, \lambda_j^*) \right\} N(\mu_j^* | m, b\lambda_j^*) W(\lambda_j^* | s, s^{-1}T) \\
& \quad \propto N\left(\mu_j^* \middle| \frac{\sum_{i \in I_j} y_i + bm}{n_j + b}, (n_j + b)\lambda_j^* \right) \\
& \qquad \times \mathcal{W}\left(\lambda_j^* \middle| s + n_j, \left[\sum_{i \in I_j} (y_i - \bar{y}_j)(y_i - \bar{y}_j)^t + \frac{n_j b}{n_j + b} (m - \bar{y}_j)(m - \bar{y}_j)^t + sT^{-1} \right]^{-1} \right),
\end{aligned}
$$

donde $\bar{y}_j$ es el promedio de las observaciones con parámetro común μ_j^*, λ_j^*, es decir, $\bar{y}_j = n_j^{-1} \sum_{i \in I_j} y_i$.

Posteriormente se actualizan los valores de $\gamma = \{m, T\}$ de acuerdo a (3.11), obteniendo, debido a la independencia de m y T (nótese que en este ejemplo no se está considerando un parámetro σ común a todas las observaciones),

$$
\begin{aligned}
p(m | \mu^*, \lambda^*, Y) & \propto N(m | a, A) \prod_{j=1}^{k} N(\mu_j^* | m, b\lambda_j^*) \\
& \propto N\left(m \middle| \left(A + b \sum_{j=1}^{k} \lambda_j^* \right)^{-1} \left(Aa + b \sum_{j=1}^{k} \lambda_j^* \mu_j^* \right), A + b \sum_{j=1}^{k} \lambda_j^* \right),
\end{aligned}
$$

y

$$
\begin{aligned}
p(T^{-1} | \lambda^*) & \propto \mathcal{W}(T^{-1} | q, q^{-1}) \prod_{j=1}^{k} \mathcal{W}(\lambda_j^* | s, s^{-1}T) \\
& \propto \mathcal{W}\left(T^{-1} \middle| q + ks, \left[s \sum_{j=1}^{k} \lambda_j^* + qI_p \right]^{-1} \right).
\end{aligned}
$$

Finalmente actualizamos el valor de α de acuerdo a lo ya mencionado en la Sección 3.3, es decir, dado un valor inicial de α y k, obtenemos un valor x con distribución $\mathcal{B}e(x | \alpha + 1, k)$ y posteriormente actualizamos a α a partir de 3.14.

Con la finalidad de probar esta algoritmo se simularon 200 observaciones de la siguiente mezcla de tres normales,

$$\frac{1}{3}N\left(\begin{pmatrix}142\\122\end{pmatrix},\Sigma\right)+\frac{1}{3}N\left(\begin{pmatrix}172\\131\end{pmatrix},\Sigma\right)+\frac{1}{3}N\left(\begin{pmatrix}200\\123\end{pmatrix},\Sigma\right),$$

donde Σ es la matriz de precisión

$$\Sigma=\begin{pmatrix}50 & 0\\ 0 & 50\end{pmatrix}^{-1}.$$

A partir de (3.6), con un calentamiento de 1000 iteraciones, se simularon $M = 2000$ observaciones de la distribución predictiva de y_{n+1} tomando una de ellas cada 10 iteraciones, y también se estimó su densidad con base en dicha expresión a partir de "promedios de Monte Carlo". Específicamente, si llamamos $(\mu^1, \lambda^1, \varphi^1, k^1, \alpha^1, (n_1^1, \ldots, n_{k^1}^1)), \ldots, (\mu^M, \lambda^M, \varphi^M, k^M, \alpha^M, (n_1^M, \ldots, n_{k^M}^M))$ a las correspondientes configuraciones y a los valores de los parámetros que dieron origen a la muestra simulada, entonces

$$f(y_{n+1}) \approx \frac{1}{M}\sum_{l=1}^{M}\left\{\frac{\alpha^l}{\alpha^l+n}N(y_{n+1}|\mu_{n+1}^l,\lambda_{n+1}^l)+\frac{1}{\alpha^l+n}\sum_{j=1}^{k^l}n_j^l N(y_{n+1}|\mu_j^l,\lambda_j^l)\right\},$$

donde $\mu_{n+1}^l, \lambda_{n+1}^l$ son datos con distribución G_0 (que a su vez depende de los valores de m y T^{-1} al momento de almacenar las muestras).

Como comentario técnico debemos señalar que las cadenas de Markov de Monte Carlo pueden quedar "atrapada" en eventos locales de la distribución final. Este es el caso en este ejemplo, pues aunque la distribución final da mayor peso a los valores $k = 3$ y $k = 4$, la probabilidad de que $k = 2$ es pequeña pero positiva, y la transición de $k = 2$ a $k = 3$ es poco probable, pues con el algoritmo implementado, sólamente una nueva observación de μ_j^*, λ_j^* puede ser agregada a la configuración (y por ende incrementarse el valor de k en 1) en cada paso (a) del algoritmo dado en la Sección 3.2, y además una vez incrementado el valor de $k = 2$ a $k = 3$ la probabilidad de regresar al estado $k = 2$ es alta, debido a que el valor de $q_{i,j}$ definido en la Sección 3.2, y que en este caso está dado en (3.22), es pequeño pues la n_j correspondiente a los nuevos valores es 1. Por otro lado, los valores $k = 2$ y $k = 3$ se alcanzan fácilmente desde la configuración $k = n$, con el que se inicializa el muestreo de Gibbs. Por esta razón, cuando implementamos el muestreo de Gibbs reinicializamos la configuración $k = n$, obteniendo los valores μ_j^*, λ_j^* para $j = 1, \ldots, n$ remuestreando de

$g_{i,0}$ dada en (3.23), sin reinicializar el valor de los hiperparámetros, en vez de correr una cadena mucho más larga.

Se analizaron las gráficas de autocorrelación de los parámetros para verificar su convergencia en el muestreo de Gibbs. Las gráficas correspondientes a la densidad predictiva de una observación futura se presentan en la Figura 3.3, donde se muestran las gráficas de contorno de la distribución verdadera de los datos, de los histogramas de la muestra original y la simulada y la estimación de la densidad a partir de la expresión anterior. Como resultado adicional del muestreo de Gibbs, obtuvimos la distribución final del número de componentes k (Tabla 3.1), donde observamos que la cantidad de modas de la distribución de la cual se generaron los datos (tres) no pertenece a la zona de mayor densidad de dicha distribución ($\{4, 5, 6, 7\}$). Lo ideal sería que el número de modas fuese la moda de la distribución final de k. Lo anterior no sucede en este ejemplo, posiblemente, debido a que el tamaño de muestra no es suficientemente grande para representar al modelo.

Cabe mencionar que el mismo ejercicio se hizo con una muestra de tamaño $n = 100$ y el modelo no alcanzó a distinguir las tres modas de la mezcla de tres normales.

Se repitió el ejercicio, pero, ahora, simulando 1000 muestras de una mezcla de tres normales con precisiones y pesos diferentes y más "encimadas",

$$\frac{2}{5}N\left(\begin{pmatrix}150\\125\end{pmatrix}, \Sigma_1\right) + \frac{2}{5}N\left(\begin{pmatrix}180\\129\end{pmatrix}, \Sigma_2\right) + \frac{1}{5}N\left(\begin{pmatrix}200\\120\end{pmatrix}, \Sigma_3\right),$$

donde,

$$\Sigma_1 = \begin{pmatrix}80 & -40\\-40 & 70\end{pmatrix}^{-1}, \qquad \Sigma_2 = \begin{pmatrix}70 & 30\\30 & 80\end{pmatrix}^{-1} \quad \text{y} \quad \Sigma_3 = \begin{pmatrix}90 & 60\\60 & 90\end{pmatrix}^{-1}.$$

Las gráficas se muestran en la Figura 3.4. En la Tabla 3.2 se despliega la distribución final del número de componentes de la mezcla, donde se observa que, en este caso, el número de modas del modelo que generó los datos sí se encuentra en la zona de mayor densidad de dicha distribución.

También, en este caso se redujo el tamaño de muestra a $n = 500$ pero la distribución predictiva de y_{n+1} no distinguió las tres modas de la mezcla.

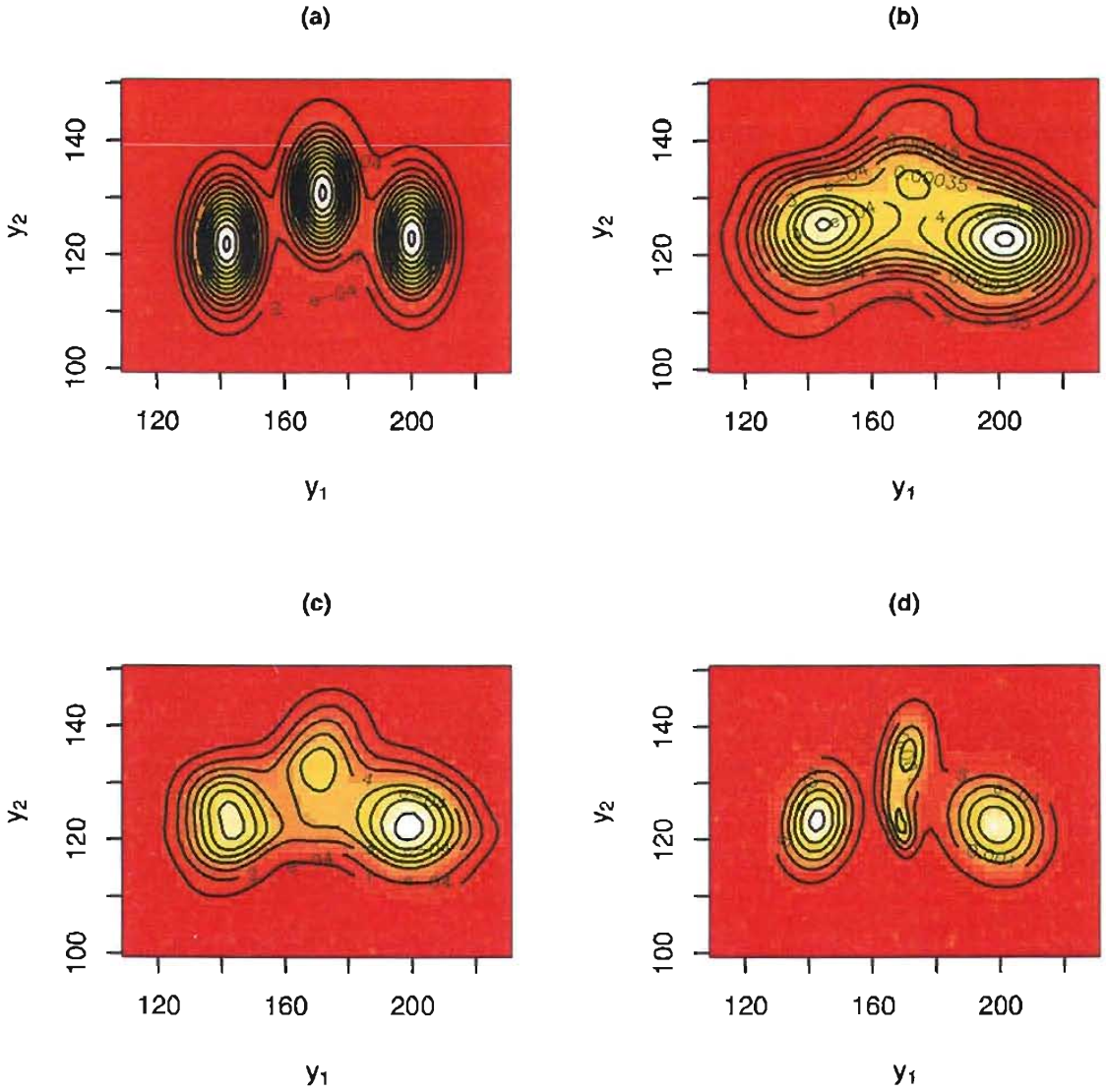

Figura 3.3. *Gráficas de contorno del modelo simulado y de las estimaciones correspondientes. (a) Densidad verdadera de los datos; (b) Densidad estimada basada en los datos originales, $n = 200$; (c) Densidad estimada basada en una muestra de la distribución predictiva final de la mezcla basada en procesos Dirichlet; (d) Estimación de Monte Carlo de la densidad predictiva final.*

Tabla 3.1. *Distribución final de k.*

k	2	3	4	5	6	7	8	9	10	11
$p(k\|Y)$	0.0010	0.0330	0.1575	0.2590	0.2620	0.1590	0.0885	0.0275	0.0070	0.0035

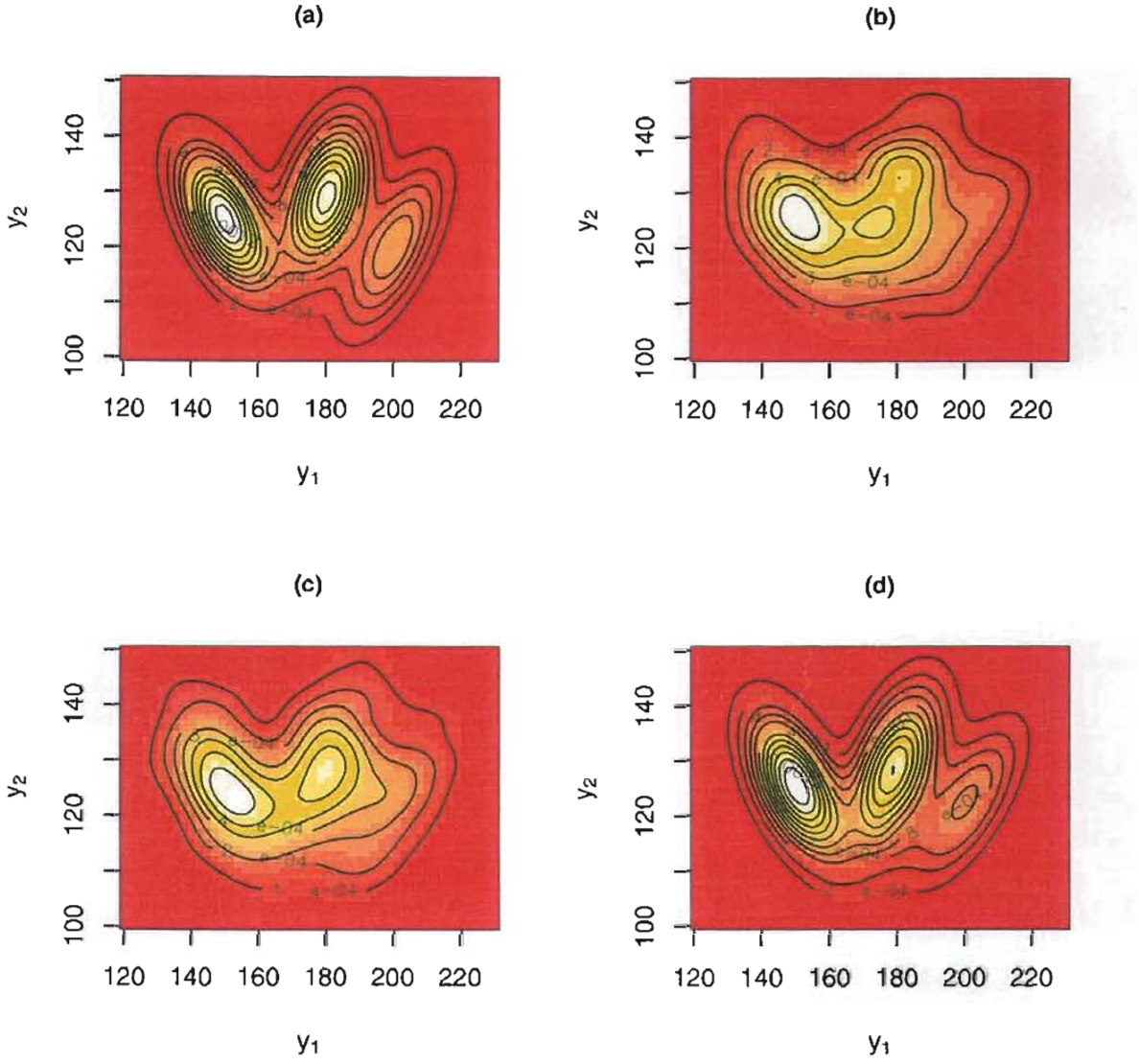

Figura 3.4. *Gráficas de contorno del modelo simulado y de las estimaciones correspondientes. (a) Densidad verdadera de los datos; (b) Densidad estimada basada en los datos originales, $n = 1000$; (c) Densidad estimada basada en una muestra de la distribución predictiva final de la mezcla basada en procesos Dirichlet; (d) Estimación de Monte Carlo de la densidad predictiva final.*

Tabla 3.2. *Distribución final de k.*

k	3	4	5	6	7	8	9	10	11
$p(k\|Y)$	0.2025	0.3060	0.2715	0.1375	0.0560	0.0170	0.0070	0.0020	0.0005

4. SELECCIÓN DE MODELOS UTILIZANDO MEZCLAS BASADAS EN PROCESOS DIRICHLET

4.1 *Planteamiento del Problema de Selección de Modelos*

Supongamos que tenemos una muestra aleatoria $Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n$ y queremos estimar la distribución de $Y_f = Y_{n+1}$. Una manera de hacerlo es utilizando métodos de estadística Bayesiana no paramétrica como el del capítulo anterior. Sin embargo, a veces deseamos un modelo sencillo que sea tratable matemáticamente y/o de fácil interpretación, lo que nos lleva a la necesidad de elegir al "mejor modelo paramétrico" de entre varios candidatos.

En general, el problema de selección de modelos se plantea de la siguiente manera: sea $\mathcal{M} = \{\mathrm{M}_\lambda : \lambda \in \Lambda\}$ una colección de modelos paramétricos, donde,

$$\mathrm{M}_\lambda = \{p_\lambda(y|\theta_\lambda), p_\lambda(\theta_\lambda)\}.$$

Es decir, M_λ es una propuesta para los modelos que inicialmente queremos ajustar a los datos. Denotaremos con f_λ a la distribución (función de densidad) final de una observación futura correspondiente a M_λ, es decir, si $y^n = \{y_1, y_2, \ldots, y_n\}$, $f_\lambda(y) \propto \int p_\lambda(y|\theta_\lambda) p_\lambda(y^n|\theta_\lambda) p_\lambda(\theta_\lambda)\, d\theta_\lambda$.

Una forma común de abordar el problema desde el punto de vista paramétrico es suponer una distribución inicial sobre $\mathcal{M}$ (y vía el Teorema de Bayes encontrar la distribución final de $\mathcal{M}$). La crítica que se hace a dicho planteamiento, Gutiérrez-Peña y Walker (2005), es que se está asignando probabilidad 1 a $\mathcal{M}$, es decir, se está asegurando que los datos verdaderamente se distribuyen de acuerdo a alguno de los elementos de $\mathcal{M}$.

Puede suceder que a la luz de los datos resulte que sea inapropiado que $\mathcal{M}$ tenga probabilidad 1 y por ende se deba agregar otro modelo a $\mathcal{M}$, y así obtener $\mathcal{M}^*$ y volver a asignar una distribución inicial a $\mathcal{M}^*$. Lo anterior es incoherente, en el sentido de que se está actualizando la distribución de los modelos candidatos sin usar el Teorema de Bayes. Por otro lado, al asignar probabilidad 1 a $\mathcal{M}$ y después verificar si es cierto (como regularmente se hace, mediante pruebas de bondad de ajuste) se está aceptando que hay incertidumbre sobre si verdaderamente los datos provienen de alguna distribución que pertenezca a $\mathcal{M}$, es decir, se está dudando que verdaderamente $\mathcal{M}$ tenga probabilidad 1.

Una manera de evitar el inconveniente expuesto anteriormente es suponer que no necesariamente la distribución de la que provienen los datos es un elemento de $\mathcal{M}$ y conformarnos con elegir el modelo en $\mathcal{M}$ que más se "aproxime" a la distribución verdadera de los datos. Replantendo el problema de selección de modelos ahora desde el punto de vista de la teoría de las decisiones, los posibles estados de la naturaleza están dados por:

$$\mathcal{F} = \{f | f \text{ es una función de densidad en Y}\}$$

De acuerdo con la teoría, requerimos asignar una distribución sobre $(\mathcal{F}, \mathcal{B})$, que como veremos más adelante será la de una mezcla basada en procesos Dirichlet, donde $\mathcal{B}$ es un σ-álgebra generado por las esferas asociadas a alguna distancia, como puede ser la distancia $\mathcal{L}_\infty$ entre funciones de densidad, pero en general, por sus propiedades matemáticas, utilizaremos la divergencia de Kullback-Leibler entre dos funciones de densidad, definida de la siguiente manera (Kullback y Leibler, 1951),

$$d_{KL}(f, g) = \int f(y) \log\left(\frac{f(y)}{g(y)}\right) dy,$$

que, si bien no es una distancia propiamente dicha (en particular no satisface la desigualdad del triángulo) es definida positiva (Kullback y Leibler, 1951). Es decir, si f y g son dos funciones de densidad, entonces $d_{KL}(f, g) \geqslant 0$, y, además, $d_{KL}(f, g) = 0$ si y sólo si $f = g$ casi seguramente.

Continuando con el planteamiento del problema desde el punto de vista de la teoría de las decisiones, necesitamos de un espacio de acciones, que en nuestro caso es Λ y así, al elegir $\lambda \in \Lambda$, estamos eligiendo al modelo M_λ. También necesitamos una función de utilidad $\mathcal{U}(\lambda, f)$ definida de $\Lambda \times \mathcal{F}$ a $\mathbb{R}$, que en nuestro caso será,

$$\mathcal{U}(\lambda, f) = \int \log\{f_\lambda(y)\} f(y)\, dy. \tag{4.1}$$

Notemos que si $\mathcal{U}(\lambda, f)$ aumenta entonces $d_{KL}(f_\lambda, f)$ disminuye y viceversa, y consideraremos como la mejor acción (el mejor modelo) a $\hat{\lambda}$ si maximiza al valor esperado de la función de utilidad dada la muestra,

$$\mathcal{U}_n(\lambda) = E(\mathcal{U}(\lambda, f)|Y^n) = \int\int \log\{f_\lambda(y)\} f(y)\, dy\, \pi(df|Y^n) = \int \log\{f_\lambda(y)\} f_n(y)\, dy, \tag{4.2}$$

donde,

$$f_n = E(f|Y^n) = \int f\,\pi(df|Y^n).$$

Equivalentemente, $\hat{\lambda}$ minimiza la divergencia esperada de Kullback-Leibler dadas las observaciones. Asimismo, utilizando que la divergencia de Kullback-Leibler es definida positiva, podemos verificar que la solución al problema de maximización de (4.2) sobre todo $\mathcal{F}$ es $f = f_n$.

Por otro lado, suponiendo que la distribución inicial de la densidad desconocida f es una mezcla basada en procesos Dirichlet, planteamos el siguiente modelo:

$$\begin{aligned}
&I. && y_i \sim F_0(y_i|\theta_i, \sigma) \text{ para } i = 1, \ldots, n.\\
&II. && \theta_1, \ldots, \theta_n \overset{i.i.d}{\sim} G.\\
&III. && G|\alpha, \gamma \sim \mathcal{DP}(\alpha G_0(\cdot|\gamma)).\\
&IV. && \alpha \sim p(\alpha),\ \gamma \sim p(\gamma) \text{ y } \sigma \sim p(\sigma).
\end{aligned}$$

Por lo que una buena aproximación de (4.2) puede obtenerse vía Monte Carlo, simulando una muestra $y_1^D, \ldots, y_M^D$ de observaciones futuras (*i.e.*, de la distribución predictiva final) a partir del algoritmo visto en el capítulo anterior y la expresión (3.6), y posteriormente calcular,

$$\mathcal{U}_n(\lambda) \approx \frac{1}{M} \sum_{i=1}^{M} \log\{f_\lambda(y_i^D)\}. \tag{4.3}$$

La ventaja de esta solución al problema de selección de modelos, sobre la paramétrica, es que si se agrega otro modelo, indexado por λ^*, a M para verificar si el nuevo modelo es preferible o mejor que los anteriores sólo debemos evaluar $\mathcal{U}_n(\lambda^*)$. De esta manera, no somos incoherentes en el sentido expuesto al principio de esta sección.

Por otro lado, al suponer un modelo basado en mezclas de procesos Dirichlet, estamos suponiendo una distribución continua para Y siempre que la F_0 que elijamos lo sea. Un ejemplo de lo aquí expuesto se presenta a continuación.

4.2 *Ejemplo de Aplicación de las mezclas basadas en Proceso Dirichlet en la Selección de Modelos*

Esta sección tiene la finalidad de ilustrar con un ejemplo bivariado y datos reales lo expuesto en la sección anterior, eligiendo de tres modelos candidatos el que mejor se ajuste a la distribución verdadera de los datos.

Los datos fueron obtenidos de un estudio de "The National Institute of Diabetes and Digestive and Kidney Diseases", en el cual se estudiaron 768 mujeres adultas de la cultura indigena Pima que vivían cerca de Phoenix. Siete variables fueron registradas, pero en este ejemplo nos avocaremos a dos y buscaremos el modelo que mejor aproxime su distribución conjunta. Ellas son: la medida de dos horas de insulina "serum" ($mu\ U/ml$) y la concentración de plasma en la glucosa en una prueba de dos horas de tolerancia a la glucosa oral. La base de datos puede encontrarse en:

http:// www.ics.uci.edu/~ mlearn/MLRepository.html.

Debemos señalar que solamente $n = 393$ de los 768 registros no tenían datos faltantes en alguna de estas variables, y como el objetivo del ejemplo es ilustrar el método de

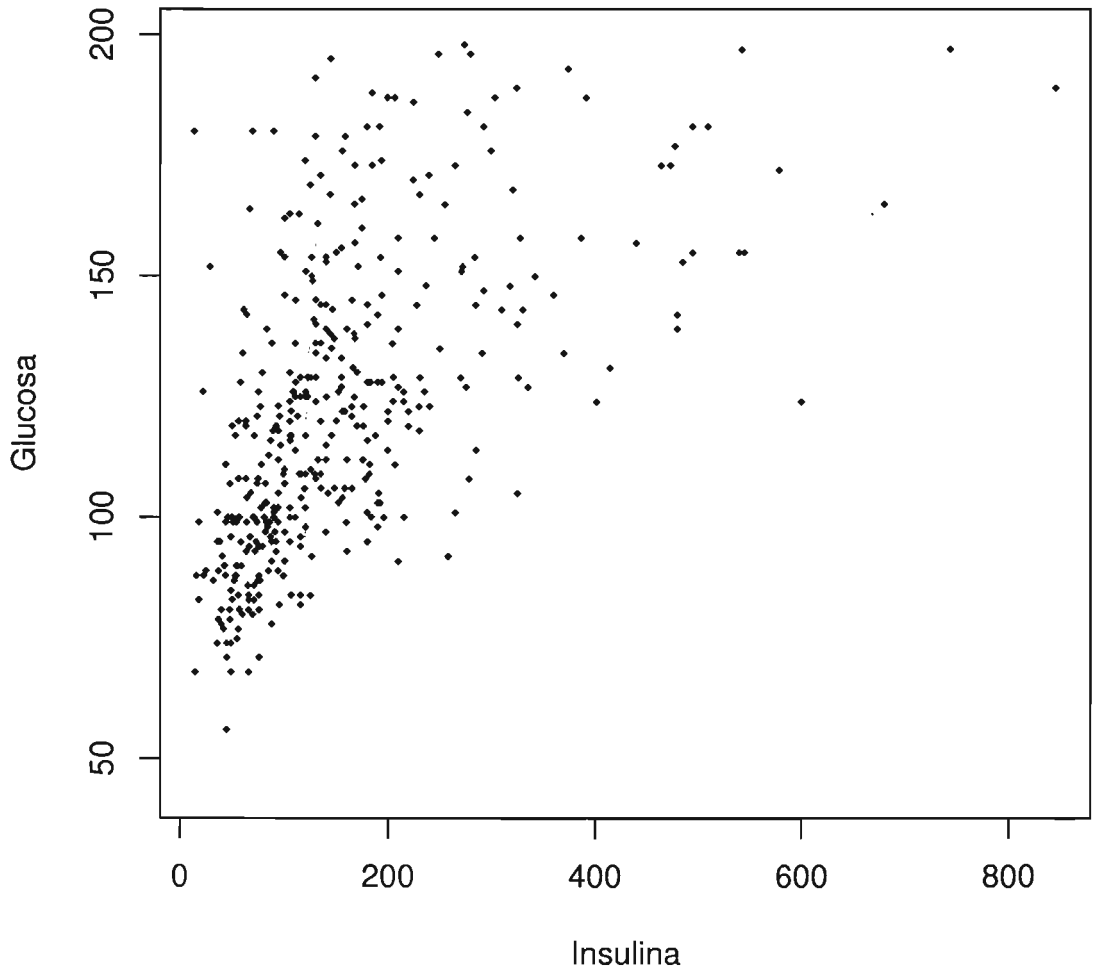

Figura 4.1 *Datos originales*

selección de modelos y no hacer un análisis estadístico de los datos, nos atrevimos a considerar únicamente a dichas observaciones. La gráfica de los datos se muestra en la Figura 4.1.

Con la finalidad de ajustar los datos a modelos sencillos pero restrictivos en la estructura de su matriz de corrrelaciones se transformaron los datos dividiendo cada observación entre la respectiva desviación estándar de la variable. Es decir, si $y_i^* = (y_{i,1}^*, y_{i,2}^*)^t$ son los valores de las dos variables de interés (en el orden en que se presentaron) del i-ésimo individuo, trabajaremos con $y_i = (y_{i,1}^*/S_1^*, y_{i,2}^*/S_2^*)$, donde $S_i^* = \sum_{j=1}^n (y_{i,j}^* - \bar{y}_i^*)^2/n$ y $\bar{y}_i^* = \sum_{j=1}^n y_{i,j}^*/n$. La gráfica de las variables transformadas se muestran en la Figura 4.2.

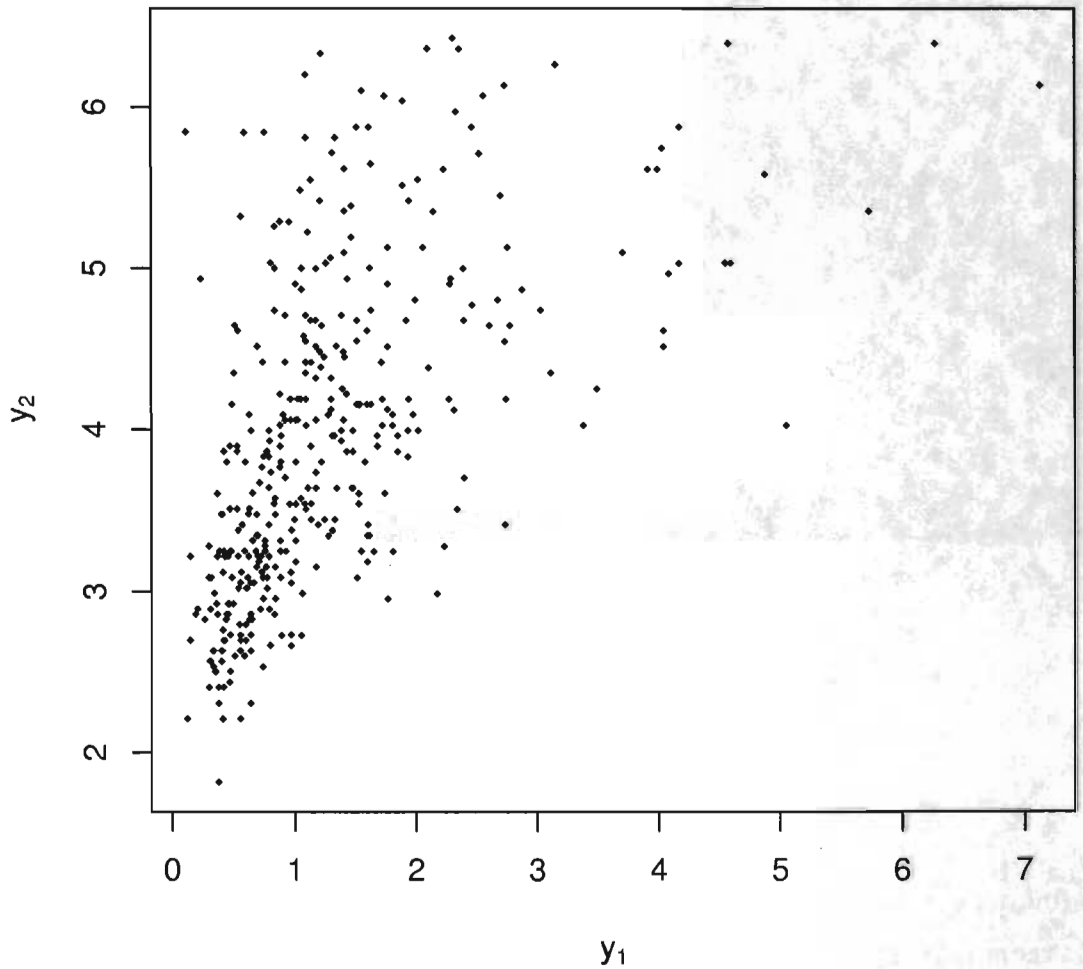

Figura 4.2 *Datos transformados*

Mediante el algoritmo dado en la Sección 3.5, se simuló una muestra de una mezcla basada en procesos Dirichlet de tamaño $M = 4000$, $y_1^D, y_2^D, \ldots, y_M^D$, con los mismos

parámetros que en el ejemplo de estimación de densidades de la Sección 3.5, que se utilizará para estimar la utilidad de los tres modelos a partir de (4.3). La gráfica de contorno de la densidad estimada de una observación futura a partir de (3.6) se presenta en la Figura 4.3.

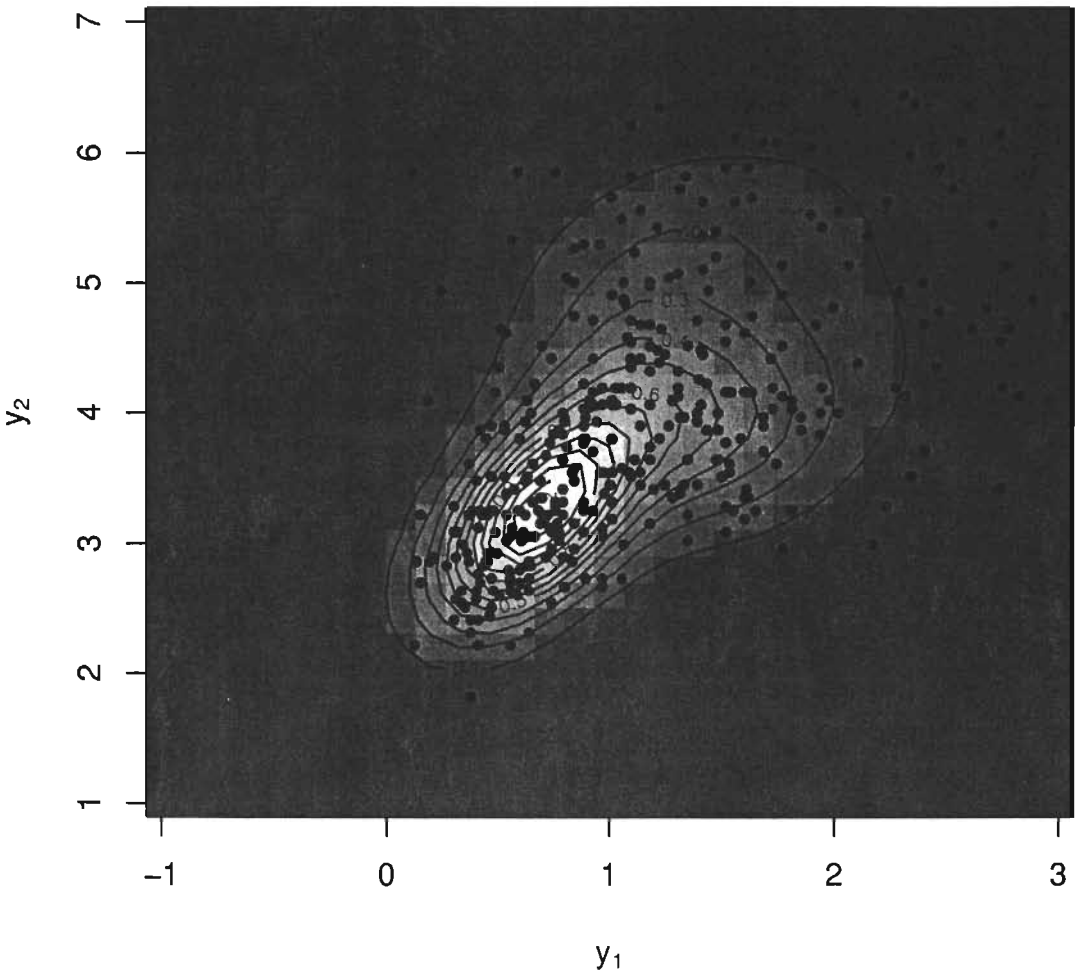

Figura 4.3 *Densidad estimada.*

Los tres modelos candidatos, buscan ser sencillos, en especial el primero que utiliza distribuciones conjugadas, y en todos se considerará una distribución inicial no informativa de los parámetros. Los modelos son los siguientes:

Primer modelo

Sea

$$y_i = (y_{i,1}, y_{i,2})^t,$$
$$y_i|\mu, \lambda \sim N(y_i|\mu, \lambda\Lambda) \text{ y}$$
$$(\mu, \lambda) \sim N(\mu|m, b\lambda\Lambda)\mathcal{G}a(\lambda|\alpha, \beta),$$

donde Λ^{-1} es la matriz de correlaciones (nótese que se está parametrizando a la distribución normal en términos de su precisión),

$$\Lambda^{-1} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}, \tag{4.4}$$

tomada de esta manera para que coincida con la matriz de correlaciones del segundo modelo y así hacer dichos modelos comparables. Los hiperparámetros m, b, α y β toman los valores de $0, 0, -1$ y 0 respectivamente, de tal forma que la distribución inicial de (μ, λ) sea no informativa, proporcional a λ^{-1}, lo que lleva a la distribución final de $y_f = y_{n+1}$ como sigue, donde p es la dimensión de y, en este caso 2,

$$p(y_f|y_1, \ldots, y_n) = Stu_p\left(y_f|m_n, \mathrm{T}, 2\alpha_n\right)$$

donde,

$$m_n = \frac{n\bar{y} + bm}{b+n},$$
$$\mathrm{T} = \frac{\alpha_n(b+n)}{\beta_n(b+n+1)}\Lambda$$
$$\alpha_n = \frac{2\alpha + pn}{2},$$
$$\beta_n = \frac{1}{2}\frac{bn}{b+n}(m-\bar{y})^t\Lambda(m-\bar{y}) + \frac{nS^2}{2} + \beta, \text{ y}$$
$$S^2 = \sum_{i=1}^{n}\frac{(y_i-\bar{y})^t\Lambda(y_i-\bar{y})}{n}.$$

La utilidad obtenida a partir de (4.3) fue de $\mathcal{U}(1) = -4.082359$ y la gráfica de contorno de la densidad de y_f dado este modelo se presenta en la Figura 4.4. Cabe mencionar que la distribución t de Student tiene 784 grados de libertad, por lo que utilizamos una densidad normal con media m_n y precisión $\frac{(2\alpha_n-2)}{2\alpha_n}\mathrm{T}$ para aproximar su densidad.

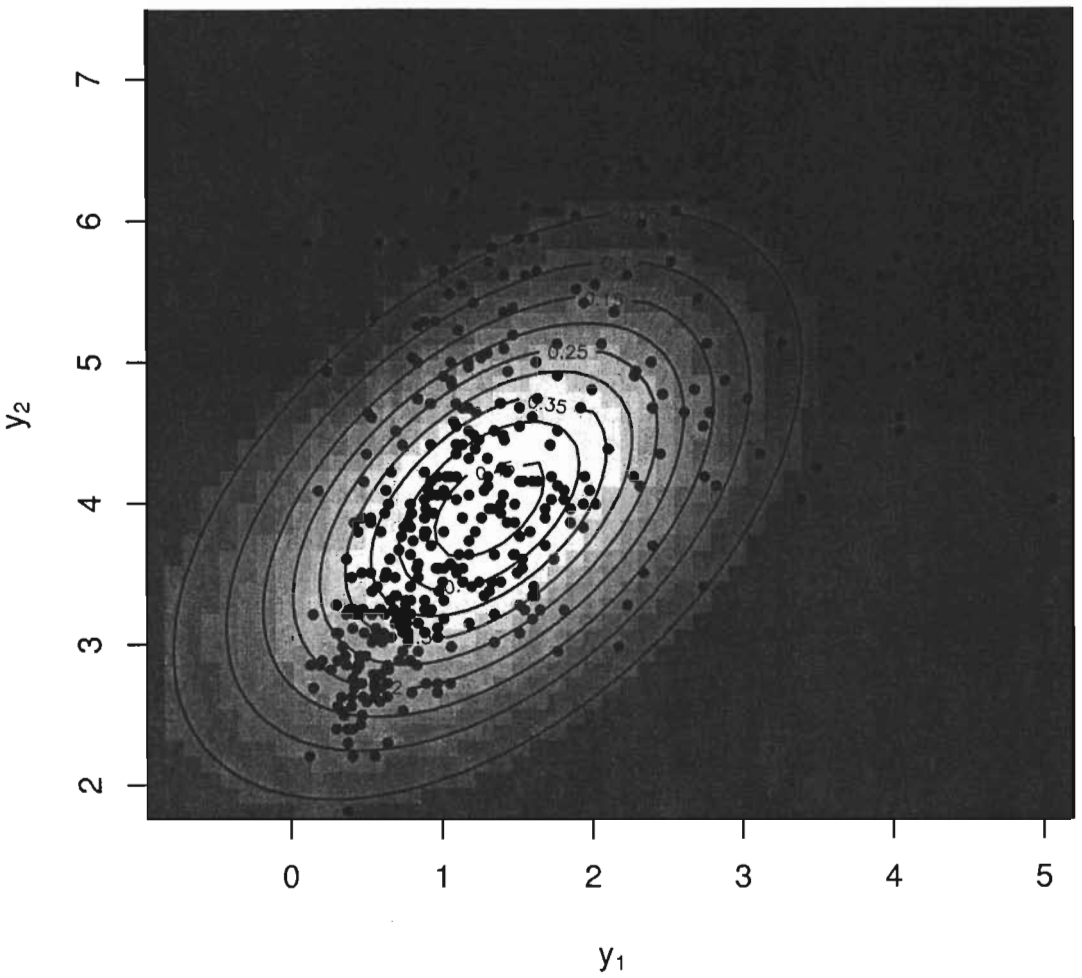

Figura 4.4 *Densidad predictiva final bajo el primer modelo.*

El segundo modelo

El segundo modelo considerado fue uno logístico:

$$y_i|\mu,\sigma \sim \text{logístico}(y|\mu,\sigma) \quad \text{y} \quad p(\mu,\sigma) = \frac{1}{\sigma},$$

donde $\mu = (\mu_1, \mu_2)^t$ la densidad del modelo logístico esta dada por,

$$p_y(y_{i,1}, y_{i,2}|\mu,\sigma) = \frac{2e^{-(y_{i,1}-\mu_1)/\sigma}e^{-(y_{i,2}-\mu_2)/\sigma}}{\sigma^2\{1+e^{-(y_{i,1}-\mu_1)/\sigma}+e^{-(y_{i,2}-\mu_2)/\sigma}\}^3}.$$

La $E(Y) = \mu$, pero este modelo es restrictivo en lo que respecta a la correlación de los datos pues su matriz de covarianzas esta dada por $\frac{\sigma^2\pi^2}{3}\Lambda^{-1}$, donde Λ^{-1} es justamente la matriz definida en (4.4), por lo que los dos modelos tienen la misma estructura de correlaciones.

Al no existir una distribución conjugada para este modelo, no es fácil obtener analíticamente la densidad predictiva de una observación futura, lo que nos lleva a la necesidad de

simular valores de μ_1, μ_2, σ dada la muestra a partir del algoritmo de muestreo-remuestreo que, dadas las densidades f y g genera variables aleatorias con distribución aproximadamente de f a partir de una muestra de g, $x_1, x_2, \ldots, x_N$, calculando,

$$q_i = \frac{w_i}{\sum_{i=1}^{N} w_i}, \qquad \text{donde,} \qquad w_i = \frac{f(x_i)}{g(x_i)},$$

y porteriormente obtenemos una muestra aproximada de f remuestreando de la densidad discreta que toma el valor x_i con probabilidad q_i †. Entre más se parezca g a f mejor será la aproximación.

Con la idea de utilizar en el papel de g a una densidad normal se hizó una reparametrización de los parámetros haciendo $\sigma = e^{\theta}$y de esta manera la densidad de y_i es,

$$p_y(y_{i,1}, y_{i,2}|\mu, \theta) = \frac{2e^{-(y_{i,1}-\mu_1)/e^{\theta}} e^{-(y_{i,2}-\mu_2)/e^{\theta}}}{e^{\theta}\{1 + e^{-(y_{i,1}-\mu_1)/e^{\theta}} + e^{-(y_{i,2}-\mu_2)/e^{\theta}}\}^3},$$

y la distribución inicial,

$$p(\mu, \theta) = 1.$$

Posteriormente calculamos distribución final de los parámetros que, por la forma de la distribución inicial, tiene la misma forma que la función de verosimilitud,

$$p(\mu, \theta|y_1, \ldots, y_n) \propto \frac{e^{-(\sum_{i=1}^{n} y_{i,1} - n\mu_1)/e^{\theta}} e^{-(\sum_{i=1}^{n} y_{i,2} - n\mu_2)/e^{\theta}}}{e^{n\theta} \prod_{i=1}^{n} \{1 + e^{-(y_{i,1}-\mu_1)/e^{\theta}} + e^{-(y_{i,2}-\mu_2)/e^{\theta}}\}^3}. \tag{4.5}$$

Encontramos numéricamente $\hat{\mu}_1, \hat{\mu}_2, \hat{\theta}$ que maximicen (4.5) y el Hessiano H, evaluado en dichos puntos, es decir,

$$H = -\frac{d}{d\mu_1, \mu_2, \theta} \log(p_y(\mu, \theta|y_1, \ldots, y_n))\Big|_{\mu_1=\hat{\mu}_1, \mu_2=\hat{\mu}_2, \theta=\hat{\theta}}. \tag{4.6}$$

Utilizamos,

$$g(\mu_1, \mu_2, \theta) = N(\mu_1, \mu_2, \theta|\hat{\mu}_1, \hat{\mu}_2, \hat{\theta}, -H/1.25) \tag{4.7}$$

para aproximar una muestra de (4.5) vía muestreo-remuestreo. Una vez obtenida dicha muestra, $(\mu_1^1, \mu_2^1, \theta^1), \ldots, (\mu_1^N, \mu_2^N, \theta^N)$, podemos estimar la densidad de una observación futura y_f dada la muestra, vía Monte Carlo, es decir,

$$p(y_f|y_1, \ldots, y_n) \approx \frac{1}{N} \sum_{i=1}^{N} p(y_f|\mu_1^i, \mu_2^i, \theta^i), \tag{4.8}$$

† Ver Rubin (1988) o Bernardo y Smith (1994), p. 351.

y utilizar (4.3) para estimar la utilidad del modelo.

En este caso, los valores de $\hat{\mu}_1, \hat{\mu}_2, \hat{\theta}$ resultaron 1.2905469, 3.9423303 y -0.1967794 respectivamente, y el negativo del hessiano,

$$-H = \begin{pmatrix} 351.0182502 & -181.86697 & 0.6956937 \\ -181.8669730 & 334.26041 & -18.6321225 \\ 0.6956937 & -18.63212 & 634.1305183 \end{pmatrix}$$

Se utilizó una muestra de g de tamaño 100,000 para obtener una muestra de la densidad final de los parámetros de tamaño 5,000. La utilidad resultante fue $\mathcal{U}(2) = -3.914865$. La gráfica de contorno de la densidad de y_f bajo este modelo se presenta en la Figura 4.5.

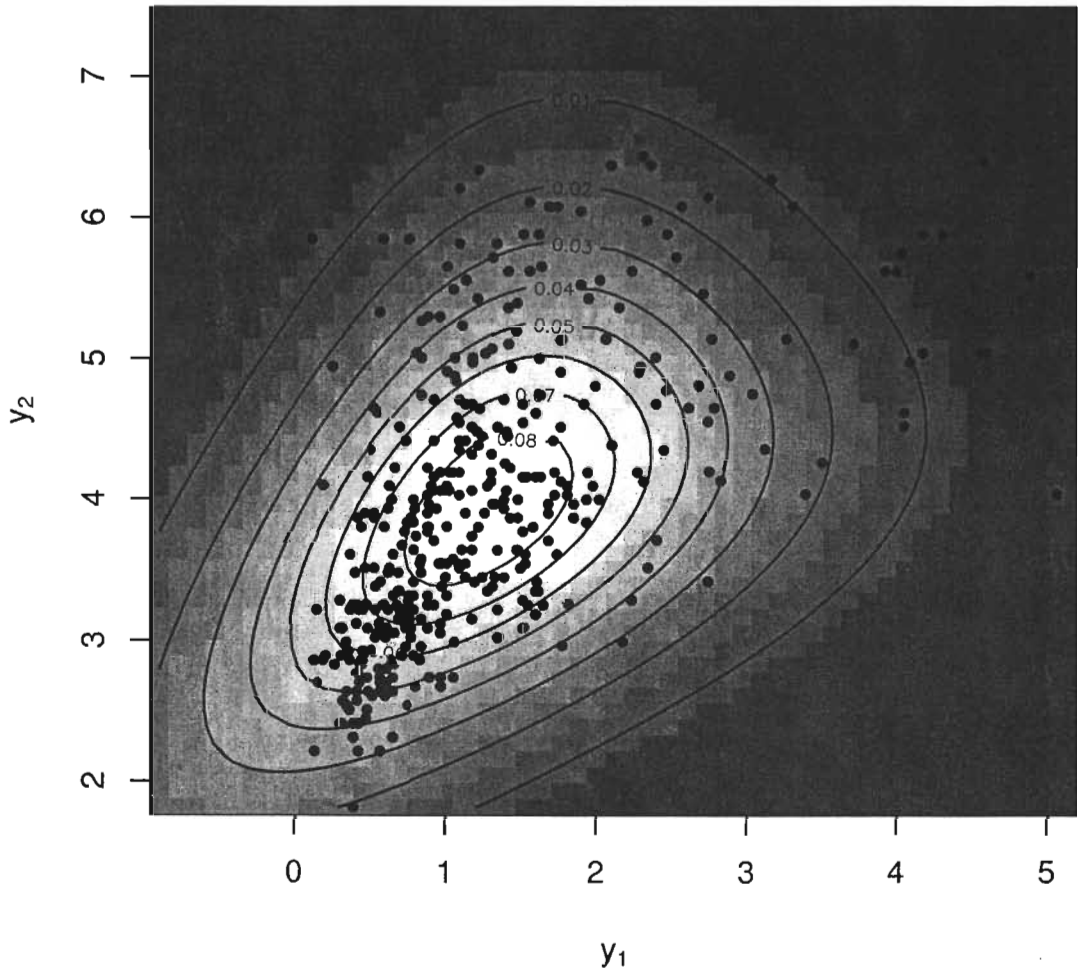

Figura 4.5 *Densidad predictiva final bajo el segundo modelo.*

El Tercer Modelo

Supongamos que

$$(U_{i,1}, U_{i,2}) \sim \mathcal{D}(\alpha_1, \alpha_2, \alpha_3),$$

y que

$$X_{i,1} = \log\left(\frac{U_{i,1}}{1 - U_{i,1} - U_{i,2}}\right) \qquad \text{y} \qquad X_{i,2} = \log\left(\frac{U_{i,2}}{1 - U_{i,1} - U_{i,2}}\right).$$

Lo que nos lleva a que,

$$p(x_1, x_2) = \frac{\Gamma(\alpha_1 + \alpha_2 + \alpha_3)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)} \frac{e^{\alpha_1 x_1} e^{\alpha_2 x_2}}{(1 + e^{x_1} + e^{x_2})^{\alpha_1+\alpha_2+\alpha_3}}.$$

Ahora, suponemos que las observaciones siguen la distribución de

$$Y_{i,1} = \mu_1 + \sigma X_{i,1} \qquad \text{y} \qquad Y_{i,2} = \mu_2 + \sigma X_{i,2}.$$

Entonces,

$$p(y_{i,1}, y_{i,2}|\mu, \sigma) = \frac{\Gamma(\alpha_1 + \alpha_2 + \alpha_3)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)} \frac{e^{\alpha_1(y_{i,1}-\mu_1)/\sigma} e^{\alpha_2(y_{i,2}-\mu_2)/\sigma}}{\sigma^2(1 + e^{(y_{i,1}-\mu_1)/\sigma} + e^{(y_{i,2}-\mu_2)/\sigma})^{\alpha_1+\alpha_2+\alpha_3}},$$

con una distribución inicial de los parámetros no informativa dada por

$$p(\mu_1, \mu_2, \sigma) = \frac{1}{\sigma}.$$

Como en el caso anterior, no es fácil encontrar la distribución predictiva final de una observación futura, por lo que para aproximar la densidad de dicha observación, vía Monte Carlo a partir de (4.8), primero necesitamos una muestra de la densidad final de los parámetros. Con el objetivo de aproximar la distribución final de los parámetros con una distribución normal para utilizar muestreo-remuestreo, reparametrizamos a $\sigma = e^{\theta}$, y de esta forma,

$$p(y_{i,1}, y_{i,2}|\mu_1, \mu_2, \theta) = \frac{\Gamma(\alpha_1 + \alpha_2 + \alpha_3)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)} \frac{e^{\alpha_1(y_{i,1}-\mu_1)/e^{\theta}} e^{\alpha_2(y_{i,2}-\mu_2)/e^{\theta}}}{e^{\theta}(1 + e^{(y_{i,1}-\mu_1)/e^{\theta}} + e^{(y_{i,2}-\mu_2)/e^{\theta}})^{\alpha_1+\alpha_2+\alpha_3}},$$

y,

$$p(\mu_1, \mu_2, \theta) = 1.$$

Por lo que la distribución final de μ_1, μ_2, θ, tiene la misma forma que la función de verosimilitud, es decir,

$$p(\mu_1, \mu_2, \theta|y_1, \ldots, y_n) \propto \frac{e^{\alpha_1(\sum_{i=1}^n y_{i,1} - n\mu_1)/e^{\theta}} e^{\alpha_2(\sum_{i=1}^n y_{i,2} - n\mu_2)/e^{\theta}}}{e^{n\theta} \prod_{i=1}^n (1 + e^{(y_{i,1}-n\mu_1)/e^{\theta}} + e^{(y_{i,2}-n\mu_2)/e^{\theta}})^{\alpha_1+\alpha_2+\alpha_3}}. \tag{4.9}$$

Posteriormente, como en el caso del segundo modelo, encontramos analíticamente, $\hat{\mu}_1, \hat{\mu}_2, \hat{\theta}$ que maximicen (4.9) y el Hessiano dado por la expresión (4.6), encontramos una muestra que se distribuye aproximadamente como (4.9) a partir del algoritmo de muestreo-remuestreo y (4.7), y de esta manera podemos calcular la densidad de y_f dados $y_1, \ldots, y_n$ a partir de (4.8), y usando (4.3) calculamos la utilidad del tercer modelo.

Utilizamos $(\alpha_1, \alpha_2, \alpha_3) = (0.4, 0.4, 0.4)$ para que la matriz de correlaciones del modelo se parezca a la matriz Λ^{-1} dada en (4.4) de manera que el modelo sea comparable con los otros dos.

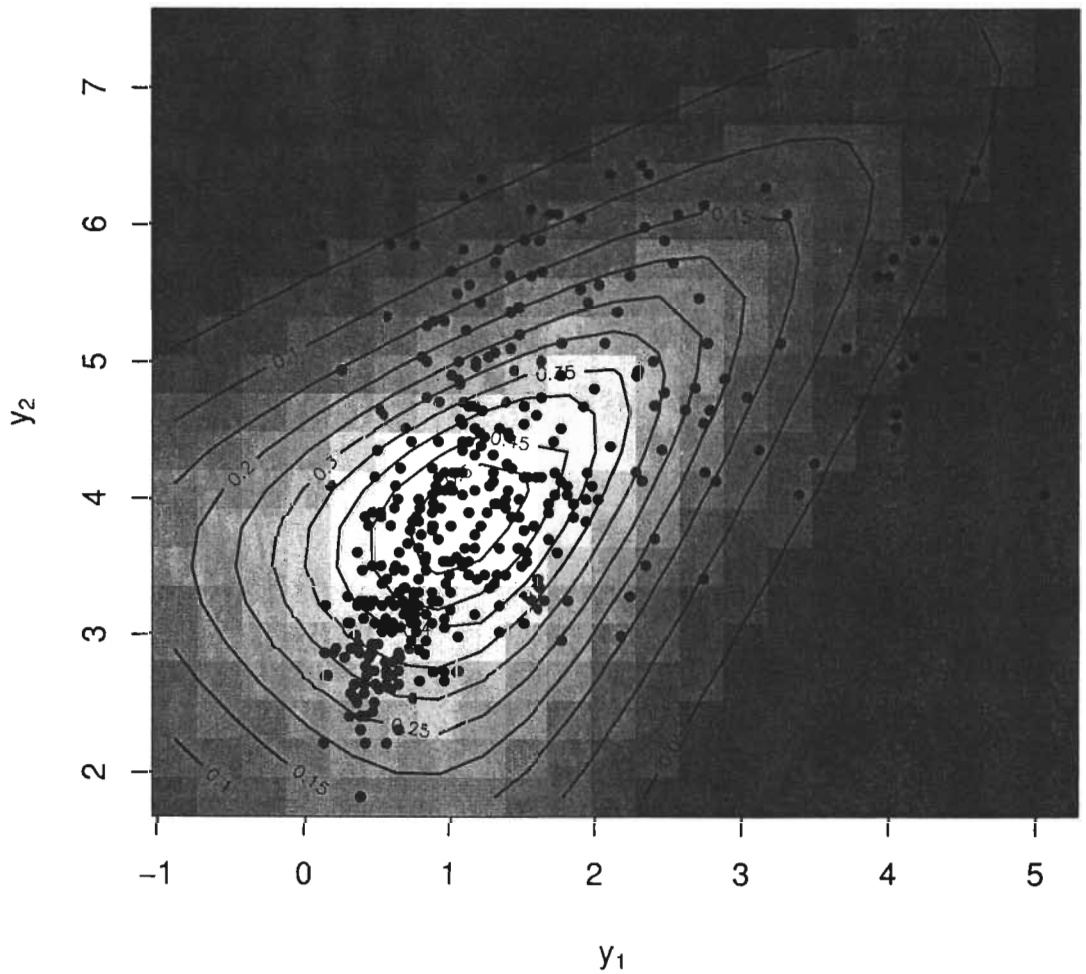

Figura 4.6 *Densidad predictiva final bajo el tercer modelo.*

En este caso, los valores de $\hat{\mu}_1, \hat{\mu}_2, \hat{\theta}$ resultaron 1.1276656, 3.8733251 y -0.7071138

respectivamente, y Hessiano multiplicado por -1,

$$-H = \begin{pmatrix} 307.72005 & -138.267532 & -29.932639 \\ -138.26753 & 271.134613 & -6.545958 \\ -29.93264 & -6.545958 & 565.182819 \end{pmatrix}$$

En el algoritmo de muestreo-remuestreo se simularon $100,000$ muestras de la densidad normal y se remuestrearon $5,000$. Obtuvimos una utilidad de $\mathcal{U}(3) = -1.958782$. La gráfica de la densidad final de y_f se muestra en la Figura 4.6.

Al haber calculado las tres utilidades, llegamos a la conclusión de que dada toda la información disponible el tercer modelo es el que mejor se aproxima a la verdadera distribución de los datos en valor esperado (de acuerdo a (4.2)). Observando las figuras, podemos notar que el tercer modelo da un mejor ajuste no sólo en la zona de mayor densidad (a diferencia del segundo), sino también en las colas (a diferencia del primero).

5. COMENTARIOS FINALES

La gran mayoría de los modelos Bayesianos utilizados actualmente en las aplicaciones, concretamente en el análisis de datos, son paramétricos. Por una parte, los modelos paramétricos son más fáciles de analizar y cuentan con técnicas más desarrolladas para su análisis. Por otro lado, muchas veces permiten describir de manera simple las características esenciales del problema estudiado y por lo tanto generalmente admiten interpretaciones más directas.

No obstante, los modelos paramétricos implican fuertes supuestos sobre el proceso que genera los datos y pueden ser poco robustos ante violaciones a dichos supuestos. Desde un punto de vista Bayesiano, suponer un modelo paramétrico implica la asignacion de probabilidad uno a un subconjunto muy pequeño dentro del conjunto de todos los modelos posibles. Lo anterior puede llevar a inferencias incoherentes si posteriormente la evidencia muestra que el modelo paramétrico no era adecuado.

Una forma de mitigar estos problemas consiste en utilizar modelos no paramétricos. Desde la perspectiva Bayesiana, esto se logra asignando distribuciones iniciales sobre espacios de funciones de densidad (o de distribución). Sin embargo, las herramientas matemáticas requeridas para el análisis de este tipo de modelos, tales como la teoría de los procesos estocasticos, son más complejas. Afortunadamente, desde hace algunos años ha sido posible analizar estos modelos a través de métodos numéricos, debido principalmente a los avances recientes en las técnicas de simulación y específicamente a las técnicas de Monte Carlo vía cadenas de Markov.

Entre los modelos Bayesianos no paramétricos propuestos hasta la fecha, las mezclas basadas en procesos Dirichlet son posiblemente los más atractivos ya que son bastantes flexibles pero al mismo tiempo son relativamente sencillos de analizar a través de las técnicas de simulación antes mencionadas.

En este trabajo hemos discutido el desarrollo y aplicaciones las mezclas basadas en procesos Dirichlet. Hemos discutido con detalle las propiedades del Proceso Dirichlet ya que forman la base de nuestro modelo no paramétrico para observaciones continuas, además de que es un modelo relevante en sí mismo y hasta la fecha es utilizado en una variedad de aplicaciones.

También hemos descrito con cierto detalle el algoritmo utilizado para obtener las muestras de la distribucion predictiva final, que son la base de nuestro análisis Bayesiano no paramétrico.

Finalmente, consideramos la aplicación de las mezclas basadas en procesos Dirichlet al problema de estimación de densidades, lo cual nos permitió posteriormente plantear y resolver un problema de selección de modelos paramétricos desde un punto de vista Bayesiano no paramétrico. Consideramos que estas ideas pueden extenderse y aplicarse a problemas similares tales como algunas generalizaciones del *bootstrap Bayesiano* (Rubin, 1981; Newton y Raftery, 1994).

Por otro lado, debemos señalar que entender al proceso Dirichlet y a las mezclas basadas en procesos Dirichlet a partir de sus referencias originales (Ferguson, 1973 y Antoniak, 1974) no nos fue sencillo, por lo que intentamos hacer más comprensibles estos conceptos unificando la notación de dichas referencias, agregando algunas demostraciones que no se incluyen en ellas por su "sencillez" y tratando de hacer más claras las otras demostraciones. Pensamos que la lectura previa, o paralela, de este trabajo a la de los artículos mencionados hace más clara y sencilla la comprensión de estos últimos.

También debemos indicar que, si bien la implementación de los modelos no paramétricos propuestos en este trabajo es un poco más complicada que la de los modelos paramétricos, las mayores dificultades prácticas de los primeros radica en la lenta convergencia del muestreo de Gibbs que se puede acelerar a partir de ciertos "trucos" (mencionados a su debido momento en este trabajo), esta dificulad también puede presentarse en los segundos siempre que se necesite llevar a cabo un muestreo de Gibbs. Por lo tanto, a nuestro parecer, es preferible utilizar técnicas no paramétricas sobre las paramétricas en el problema de estimación de densidades, ya que, aunque los cálculos necesarios en la estimación de densidads a partir de modelos no paramétricos son un poco más complicados, la ganacia la obtenemos al conseguir estimaciones más robustas. Cabe mencionar que aunque las mezclas basadas en procesos Dirichlet es continua con probabilidad 1, esto no nos asegura que no haya un subconjunto propio de las distribuciones continuas con probabilidad 1.

Además, si no es práctico modelar la distribución de los datos a partir de mezclas basadas en procesos Dirichlet, este modelo nos da un criterio que permite seleccionar el

modelo que mejor se aproxime a la verdadera distribución de los datos en valor esperado dada la muestra de una gama de modelos candidatos.

REFERENCIAS

Abramowitz, M. y Stegun, I.A. (1964). *Handbook Mathematical Functions.* National Bureau of Standards.

Antoniak, C.E. (1974). Mixture of Dirichlet processes with applications to nonparametric problems. *Annals of Statistics*,**2**, 1152-74.

Bernardo, J.M. y Smith, A.F.M (1994). *Bayesian Theory.* John Wiley & Sons, New York.

Blackwell, D. y MacQueen, J.B. (1973). Ferguson distributions via Pólya schemes. *Annals of Statistics*,**1**, 353-355.

Dey, D., Sinha, D. y Müller, P. (1998). *Practical nonparametric and seminonparametric Bayesian statistics.* Lecture Notes in Statistics, Vol. 133. Springer-Verlag, New York.

Dubins, L.E. y Freedman, D.A. (1965). Random distribution functions. *Bulletin of the American Mathematical Society*, **69**, 548-551.

Dubins, L.E. y Freedman, D.A. (1967). Random distribution functions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, **2**, 183-214. Univ. of California Press.

Escobar, M.D. (1988). Estimating the means of several normal poplulations by nonparametric estimation of the distribution of the means. Unpublished Ph.D. dissertation, Department of Statistics, Yale University.

Escobar, M.D. y West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90**, 577-588.

Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, **1**, 209-230.

Ferguson, T.S. (1974). Prior distributions on spaces of probability measures. *Annals of Statistics*, **2**, 615-629.

Freedman, D.A. (1963). On the asymptotic behavior of Bayes estimates in the discrete case I. *Annals of Mathematical Statistics*, **34**, 1386-1403.

Freedman, D.A. (1965). On the asymptotic behavior of Bayes estimates in the discrete case II. *Annals of Mathematical Statistics*, **36**, 454-456.

Gelfand, A.E. y Smith, A.F.M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398-409.

Gutiérrez-Peña, E. y Walker, S.G. (2005). Statistical decision problems and Bayesian nonparametric methods. Aceptado para su publicación en el *International Statistical Review.*

Korwar, R. y Hollander, M. (1973). Contributions to the theory of Dirichlet processes. *Annals of Probability*, **1**, 705-711.

Kingman, J.F.C. (1975). Random discrete distributions. *Journal of the Royal Statistical Society B*, **37**, 1-22.

Kraft, C.H. (1964). A class of distribution function processes which have derivatives. *Journal of Applied Probability*, **1**, 385-388.

Kraft, C.H. y Vann Eeden, C. (1964). Bayesian bio-assay. *Annals of Mathematical Statistics*, **35**, 886-890.

Krasker, W. y Pratt, J.W. (1986). Discussion of "On the consistency of Bayes estimates" by Diaconis and Freedman. *Annals of Statistics*, **14**, 55-58.

Kolmogorov, A.N. (1933). *Foundation of the Theory of Probability.* Segunda edición. Chelsea, New York.

Kullback, S. y Leibler, R.A. (1951). On information and sufficiency. *Annals of Statistics*, **22**, 79-86.

Lavine, M. (1992). Some aspects of Pólya tree distributions for statistical modelling. *Annals of Statistics*, **20**, 1203-1221.

Lavine, M. (1994). More aspects of Pólya trees for statistical modelling. *Annals of Statistics*, **22**, 1161-1176.

Lo, A.Y. (1984). On a class of Bayesian nonparametric estimates I. Density estimates. *Annals of Statistics*, **12**, 351-357.

MacEachern, S.N. (1994). Estimating normal means wih a conjugate style Dirichlet process models. *Comunications in Statistics: Simulation and Computation*, **23**, 727-741.

MacEachern, S.N. y Müller, P. (1998). Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*, **7**, 233-238.

Newton, M.A. y Raftery, A.E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society B*, **56**, 3-48 (con discusión).

Rubin, D.B. (1981). The Bayesian bootstrap. *Annals of Statistics*, **9**, 130-134.

Rubin, D.B. (1988). Using the SIR algorithm to simulate posterior distributions. *In Bayesian Statistics*, **3** (Editores: J.M. Bernardo et al). University Press, Oxford, 395-402.

Schervish, M.J. (1995). *Theory of Statistics*. Springer-Verlag, New York.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistaca Sinica*, **4**, 639-650.

Walker, S.G., Damien, P., Laud, P.W. y Smith, A.F.M. (1999). Bayesian nonparametric inference for random distributions and related functions. *Journal of the Royal Statistical Society B*, **61**, 485-527 (con discusión).

West, M. (1990). Bayesian kernel density estimation. *ISDS Discussion Paper 90-A02*, Duke University.

West, M. (1992). Hyperparameter estimation in Dirichlet Process mixture models. *ISDS Discussion Paper 92-A03*, Duke University.

West, M., Müller, P. y Escobar M.D. (1994). Hierarchical priors and mixture models, with applications in regression and density estimation. *Aspects of Uncertainty: A Tribute to D. V. Lindley* (editores: AFM Smith y PR Freeman). Wiley, New York, 363-386.

Wilks S.S (1962). *Mathematical Statistics*. Wiley, New York.